

THE LIBRARY

The Ontario Institute
for Studies in Education

Toronto, Canada





Digitized for Microsoft Corporation
by the Internet Archive in 2008.

From University of Toronto.

May be used for non-commercial, personal, research,
or educational purposes, or any fair use.

May not be indexed in a commercial service.

The Journal of Educational Psychology

*Devoted Primarily to the Scientific Study of Problems of
Learning and Teaching*

BOARD OF EDITORS:

HAROLD ORDWAY RUGG, *Chairman.*
Lincoln School of Teachers College.
Teachers College, Columbia University.

RUDOLF PINTNER,
Teachers College, Columbia University.

JAMES CARLETON BELL,
Brooklyn Training School for Teachers.

BEARDSLEY RUML,
Carnegie Corporation, New York City.

FRANK NUGENT FREEMAN,
University of Chicago.

LEWIS MADISON Terman,
Leland Stanford University.

ARTHUR IRVING GATES,
Teachers College, Columbia University.

EDWARD LEE THORNDIKE,
Teachers College, Columbia University.

VIVIAN ALLEN CHARLES HENMON,
University of Wisconsin.

LAURA ZIRBES, *Assistant Editor.*
Lincoln School of Teachers College.



Published Monthly Except June to August by
WARWICK and YORK, Inc.,
York, Pa. Baltimore, Md.

LIST OF CONTENTS VOLUME XII

J. CARLETON BELL. <i>Group Tests of Intelligence: An Annotated List</i>	103
B. R. BUCKINGHAM. <i>Intelligence and its Measurement: A Symposium</i>	271
G. T. BUSWELL. <i>The Relationship Between Eye Perception and Voice Response in Reading</i>	217
FOWLER D. BROOKS. <i>Rate of Mental Growth, Ages Nine to Fifteen</i>	501
CLARA F. and LAURA M. CHASELL. <i>A Survey of the Three First Grades of the Horace Mann School by Means of Psychological Tests and Teachers' Estimates, and a Statistical Evaluation of the Measures Employed</i>	72, 243
CECILE COLLOTON and HAROLD RUGG. <i>Constancy of the Stanford-Binet I.Q. as shown by Retests</i>	315
S. S. COLVIN. <i>Intelligence and its Measurement: A Symposium</i>	136
W. F. DEARBORN. <i>Intelligence and its Measurement: A Symposium</i>	211
F. N. FREEMAN. <i>The Interpretation and Application of the Intelligent Quotient</i>	3
<i>Intelligence and its Measurement: A Symposium</i>	133
<i>Comments on Professor Peterson's Criticism</i>	155
<i>The Scientific Evidence on the Handwriting Movement</i>	253
RAYMOND FRANZEN and F. B. KNIGHT. <i>Criteria to Employ in Choice of Tests</i>	408
ARTHUR I. GATES. <i>Educational Psychology at the Chicago Meeting of Scientific Societies</i>	63
<i>The True-False Test as a Measure of Achievement in College Courses</i>	276
<i>An Experimental and Statistical Study of Reading and Reading Tests</i>	303, 378, 445
ARTHUR I. GATES, CHAS. H. JUDD and LAURA ZIRBES. <i>Special Review of Mrs. Burgess' Monograph on Silent Reading</i>	347
M. E. HAGGERTY. <i>Intelligence and Its Measurement: A Symposium</i>	212
V. A. C. HENMON. <i>An Experimental Study of the Value of Word Study</i>	98
<i>Intelligence and its Measurement: A Symposium</i>	195
J. P. HERRING. <i>Verbal and Abstract Elements in Intelligence Examinations</i>	511
CHAS. H. JUDD, ARTHUR I. GATES and LAURA ZIRBES. <i>Special Review of Mrs. Burgess' Monograph on Silent Reading</i>	347
T. L. KELLEY. <i>Transmutation of Values on the Thorndike and Ayres Handwriting Scales: A Correction</i>	288
F. B. KNIGHT and RAYMOND FRANZEN. <i>Criteria to Employ in Choice of Tests</i>	408
E. E. LINDSAY. <i>Personal Judgments</i>	413

HELEN MARSHALL AND RUDOLF PINTNER. <i>A Combined Mental-Educational Survey.</i>	23
<i>Results of the Combined Mental-Educational Survey Tests.</i>	82
LEIGH MUDGE. <i>Time and Accuracy as Related to Mental Tests</i>	159
GARRY C. MYERS. <i>Prophecy of Learning Progress by Beta</i>	228
L. A. PECKSTEIN. <i>Masses vs. Distributed Effort in Learning.</i>	92
JOSEPH PETERSON. <i>The Growth of Intelligence and the Intelligence Quotient</i>	148
<i>Intelligence and its Measurement: A Symposium</i>	98
RUDOLF PINTNER AND HELEN MARSHALL. <i>A Combined Mental-Educational Survey.</i>	32
<i>Results of the Combined Mental-Educational Survey Tests.</i>	82
RUDOLF PINTNER. <i>Intelligence and its Measurement: A Symposium</i>	139
LOUISE E. POULL. <i>Constancy of I.Q. in Mental Defective, According to the Stanford Revision of Binet Tests.</i>	323
L. W. AND S. L. PRESSY. <i>A Critical Study of the Concept of Silent Reading Ability</i>	25
S. L. PRESSY. <i>Intelligence and its Measurement: A Symposium</i>	144
HAROLD RUGG AND CECILE COLLOTON. <i>Constancy of the Stanford-Binet I.Q. as shown by Retests</i>	315
HAROLD RUGG. <i>Is the Rating of Human Character Practicable?</i>	425, 485
B. RUMML. <i>Intelligence and Its Measurement: A Symposium.</i>	143
LEWIS M. TERMAN. <i>Mental Growth and the I.Q.</i>	325, 401
<i>Intelligence and Its Measurement: A Symposium.</i>	127
PAUL W. TERRY. <i>The Reading Problem in Arithmetic</i>	365
E. L. THORNDIKE. <i>The Constitution of Arithmetical Abilities</i>	14
<i>Intelligence and its Measurement: A Symposium</i>	124
<i>The Psychology of Drill in Arithmetic: The Amount of Practice.</i>	183
L. L. THURSTONE. <i>Intelligence and its Measurement: A Symposium.</i>	202
HERBERT WOODROW. <i>Intelligence and its Measurement: A Symposium.</i>	207
J. E. WALLACE WALLINS. <i>The Results of Retests by Means of the Binet Scale</i>	392
ANGELINA WEEKS. <i>The Terman Vocabulary as a Group Test.</i>	531
J. WYMAN AND MIRIAM WENDLE. <i>What is Reading Ability?</i>	517
LAURA ZIRBES, ARTHUR I. GATES AND CHAS. H. JUDD. <i>Special Review of Mrs. Burgess' Monograph on Silent Reading</i>	347

THE JOURNAL OF EDUCATIONAL PSYCHOLOGY

Vol. XII

JANUARY, 1921

No. 1

ANNOUNCEMENT OF THE REORGANIZATION OF THE JOURNAL OF EDUCATIONAL PSYCHOLOGY.

With this issue the Editorial Board of the JOURNAL OF EDUCATIONAL PSYCHOLOGY is reorganized and enlarged. The new board is now engaged in formulating the policy of the Journal and in outlining its contents for the year 1921. The details of this policy and of the outline of contents will be announced in the next issue. We wish now to comment briefly upon the broader outline of that policy.

The Journal will be devoted primarily to the scientific study of problems of learning and teaching. The editors' chief purpose is to make it a clearing house for the discussion of scientific investigation and experimentation, which is necessary to improve permanently teaching in the public schools. They will attempt to fulfill this purpose by organizing series of articles, by creating discussion departments, and by making the Journal a prompt and complete clearing house for the review of pertinent educational publications in the field. To do that they are now securing series of articles by leading workers in educational psychology and educational experimentation on such matters as:

1. Desirable changes in public school curricula from the standpoint of experimentation and investigation of learning. Professor Thorndike's article on "The Constitution of Arithmetical Abilities" in the present issue is the first article of this series. It will be followed by others written by leaders who have investigated the learning and teaching of writing, spelling, reading, history, geography, civics, home economics and high school mathematics. Thus the Journal should appeal to public school supervisors and pro-

gressive teachers, as well as to specialists in educational psychology.

2. Critical and evaluative articles on the construction and use of intelligence and educational tests. The JOURNAL OF EDUCATIONAL PSYCHOLOGY was one of the first publication agencies to commit itself definitely to the support and improvement of the intelligence test movement. The editors have an important interest in the continuation of that leadership by the Journal.

3. Articles and discussions of the content of courses in educational psychology which will appeal to normal school teachers of the subject. The editors wish to contribute as much as possible to the improvement of the training of public school teachers and believe that the application to normal school curricula of current scientific thinking in educational psychology will go far towards bringing about this result.

4. A new department for discussion of research problems has been organized, and will be conducted by Miss Laura Zirbes. The first publication of material from this department appears in the present issue.

5. Each month the Journal will attempt to acquaint its readers with the articles in educational psychology that have appeared in the current issue of other magazines by means of brief annotated references.

6. The editors wish to stimulate the publication of important miscellaneous findings from investigation and experimentation. Beginning with the February issue, they will present, in short discussions, statistical findings and concise interpretations from investigations of learning, intelligence and educational testing, correlation work and the like.

7. The editors will make a serious attempt to review in each issue of the Journal books, monographs and other types of important contribution in educational psychology and closely related fields of education which have appeared within the previous month or two.

THE INTERPRETATION AND APPLICATION OF THE INTELLIGENCE QUOTIENT

FRANK N. FREEMAN,
University of Chicago.

The purpose of this paper is to discuss the relationship between the IQ as a measure of the mental capacity of the individual and the facts of mental development. The IQ was first used as a means of expressing the relative mental ability of individuals as measured by the Binet scale. The first measure which was used with this scale was *mental age*. This, however, is an absolute and not a relative measure. The first relative measure used was the difference between the mental age and the chronological age, which may be called the *mental age difference*. This expresses the individual's superiority or inferiority in terms of a year's mental growth as a unit. It was soon discovered, however, that this measure, when applied to the Binet scale, would result in an apparent greater retardation or acceleration in the case of older children than in the case of younger ones.

The fundamental explanation of this fact was found in the greater overlapping of the scores of older children than of younger ones. To give a specific illustration, in the case of a number of tests which were used in the Binet scale and which were found standard for the fifth year, about twenty-three per cent more six year old than five year old children passed the tests. When the comparison was made with the tests for ten year olds, however, it was found that not until the twelfth year was reached did the older group excel the younger group by twenty-three per cent. Another way of putting the matter is to say that the five year old group overlapped the six year old group to the same degree as the ten year old group overlapped the twelve year old group. As a consequence of this fact, a six year old pupil whose score equaled that of the median of the five year old child was found to occupy about the same position in the distribution of his group as was occupied by the twelve year old child whose score equaled that of a ten year old. The child's mental capacity could be represented by a ratio between his mental age and his chronological age and this ratio would remain constant. To put it in another way, the use of such a ratio gives a distribution of indices of ability which have approximately the

same range from year to year. This makes the measures comparable throughout the period of the individual's mental growth.

We have been dealing thus far with facts of observation and with their use empirically to get measures of ability which shall be comparable from year to year and which will fit a particular scale. Thus far, no theory of mental development has been proposed. The point with which we are next concerned is the explanation of the fact that the IQ is the measure of ability which fits the Binet scale, and following this will come the question of the application of these facts and the consequent theories to the methods of expressing the results of other scales. To be specific, the IQ, or the ratio of mental age to chronological age, has been used in connection with certain point scales. The justification for this use will be one of the subjects of our inquiry.

It may be noted, parenthetically, that we are not dealing here with the ratio used by Yerkes in his point scale, namely, the coefficient of intellectual ability. This coefficient is calculated by finding the ratio between the individual's score and the median score for his age. We may, however, in connection with the other discussion, consider the implication of this index also.

What now is the nature of intellectual ability which is implied by the use of the IQ? The most common assumption concerning this matter is that the rate of intellectual growth is not uniform but regularly decreases with advancing age. The curve of mental growth which may be taken to satisfy roughly the requirement of the IQ is given in Woodrow's "Brightness and Dullness of Children." See Figure 1. This curve is apparently drawn somewhat empirically. The form of the curve is not in this case the sole condition of the validity of the IQ, but it is the chief one. In addition to a decreasing rate of mental growth we find here represented also some divergence between the lines of growth which represent different levels of ability.

In order to illustrate the way in which the IQ works out at various ages, and remains constant at successive ages for a given level of ability, four cases of individuals on the moron level have been represented at the chronological ages 3, 6, 9 and 15. The mental ages are indicated by the points on the normal curve at which it is cut by the horizontal lines drawn from the points on the curve for the moron

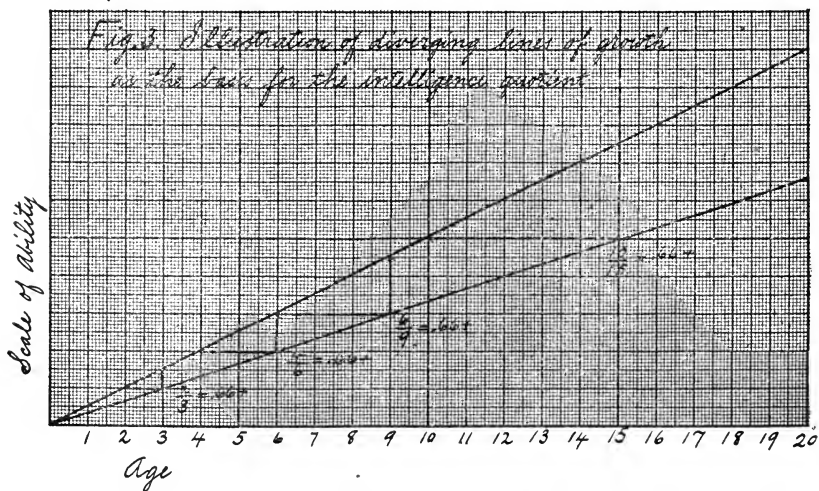
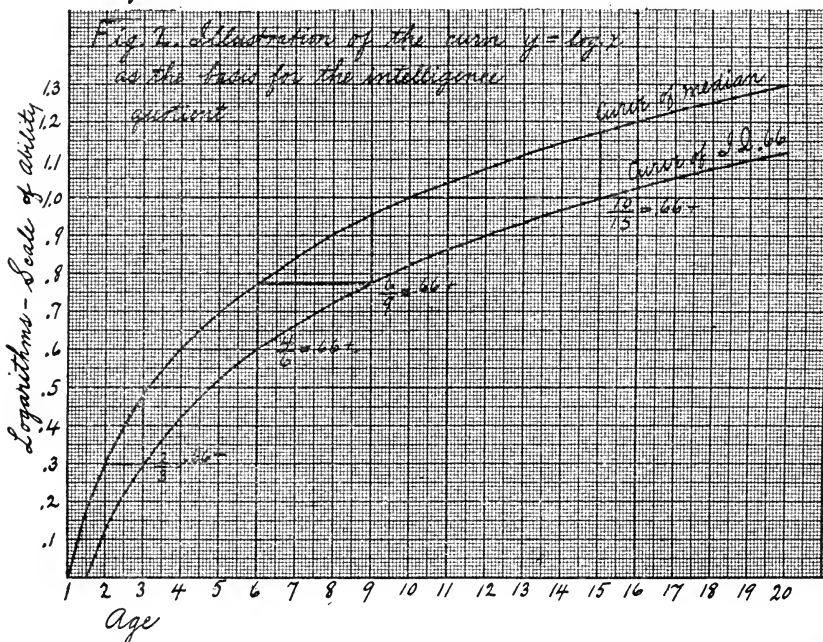
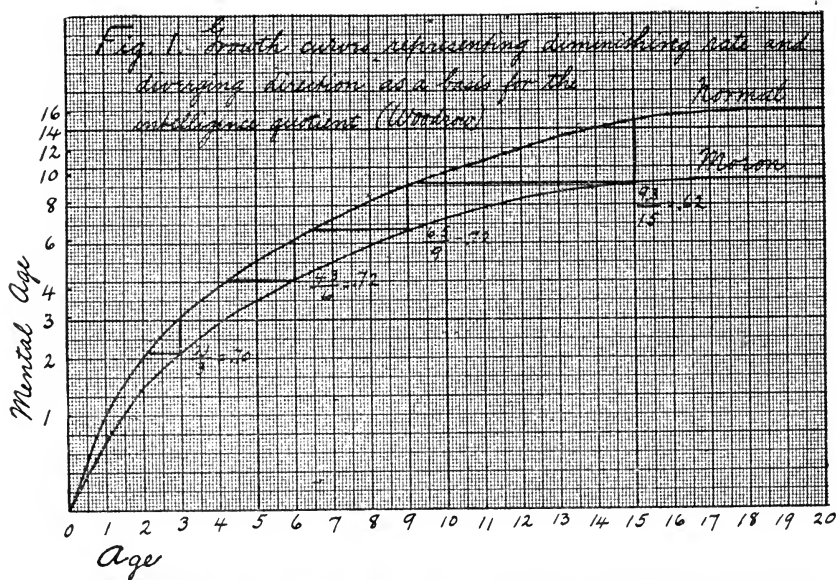
level which correspond to the respective chronological ages. The divergence of the two lines is indicated by the increasing length of the vertical lines which join them at the respective ages. The decrease in the IQ at age 15 indicates that the lines diverge too much at this part or that they approach the horizontal too quickly.

It may be shown mathematically that if the rate of growth were the only factor determining the IQ, and if the IQ were valid, the curve would be logarithmic, the formula being $y = \log x$. This is illustrated in Figure 2. Assuming that the growth curves of all individuals, including superior and inferior ones, followed this form and were equidistant from one another on the vertical axis, this form of development would produce the degree of overlapping in successive years implied in the IQ. Compare for example the ratio of mental age to chronological age in the case of illustrative individuals of the chronological ages, 3, 6, 9 and 15, and mental ages, 2, 4, 6 and 10, respectively.

This, however, is only one of two possible conditions which may alone produce the increasing overlapping necessary to make valid the IQ. The second possible factor is illustrated in Figure 3. We have here represented a form of mental growth which follows a straight line. The lines of development of different individuals, however, do not run parallel to one another, but diverge. Thus the range of distribution at successive years increases proportionately to the size of the score. The way in which this supports the validity of the IQ is illustrated again by the relation of mental age to chronological age of inferior children at the ages of 3, 6, 9 and 15. These two factors might, of course, be combined in various proportions.

The Binet scale is so constituted that it gives no indication whether the one or the other of these two assumptions is the correct one or whether there exists in fact a combination of them. As already remarked, Woodrow assumes some divergence on growth curves as well as a decrease in rate of growth. The Binet scale does not give an individual's intelligence rating in terms of a scale which is independent of the age levels themselves. Mental rating is always in terms of abilities at different age levels and therefore it is impossible to determine the nature of growth except in age terms.

In order to get a further line upon this problem, it is necessary to examine the age norms in those tests which use the same scale



throughout the ages which are to be compared. This is true of the various point scales which are now in common use.

We may first inquire whether the curve of mental growth follows or approaches the logarithmic curve, or in more general terms what its form is. We have been given age norms for several point scales. The validity of these norms may be questioned beyond the age of thirteen or fourteen, on the ground of a selection of cases beyond these years. We are safe, therefore, only in reference to the years below this.

The first point scale for which we have norms is the one devised by Yerkes. The curve is shown in Figure 4. A casual glance at this figure might give the impression that it represents a decreasing rate of growth. More careful inspection, however, shows that it is made up of two parts, one for the years four to twelve, and the other for the later years. The line of the norms from four to twelve follows almost exactly a straight line. Since this is the most reliable part of the curve, we need not consider the rest.

The second growth curve, based on a large number of children tested by a point scale, is that of Pressey, as shown in Figure 5. This gives very nearly a straight line curve also, extending to age sixteen, although there is a slight divergence covering years eleven to fourteen. If we extend the examination only to years eight to fourteen, the line of development is practically straight. Pressey gives also the norms for this cross-out test scale, beginning with age four. The curve for these is shown in Figure 6. This gives practically a straight line growth up to and including age nine. At ages ten and eleven we are apparently approaching, as in many other cases, the limits of difficulty of the tests.

Figure 7 shows the same facts for the Otis Test, Figure 8 for the Haggerty Tests, Figure 9 for the Pintner Non-language Tests, Figure 10 for the National Intelligence Test, and Figure 11 for the Illinois Examination. In all of these but the Pintner and the National Intelligence Test, the line of growth is almost exactly straight. In fact, the norms which Otis derives from his data, represented by the light lines in Figure 7, follow exactly a straight line up to age fourteen. This may be compared with the actual growth curves which are represented by the heavy lines and which were derived from the table of the distribution scores.

It is apparent that the assumption that intellectual growth follows a curve which approaches the logarithmic curve is not borne

Fig. 8. Norms for Haggerty's Delta 1 and Delta 2

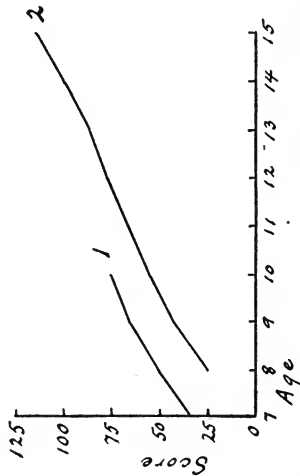


Fig. 10. Provisional age standards, National Intelligence Tests

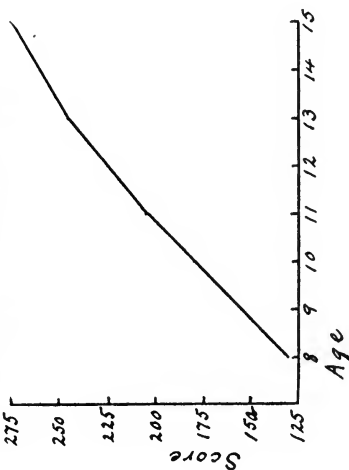


Fig. 9. Curves for three percentile groups in Pintner's non-language Test

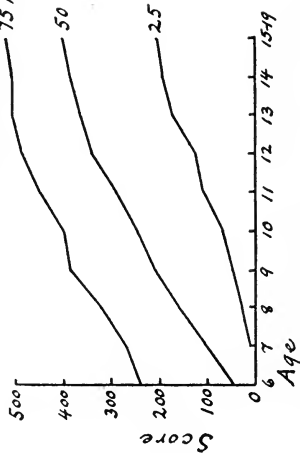
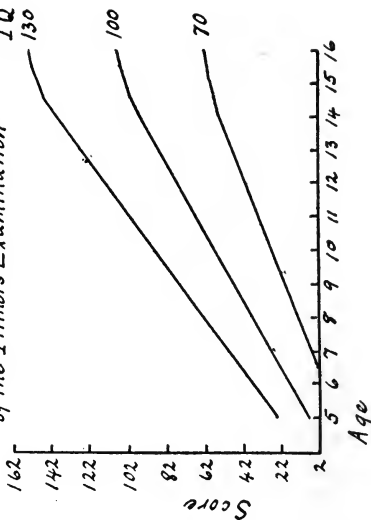


Fig. 11. Three IQ age lines from norms of the Illinois Examination IQ



out by the results of the point scale examinations. There are two conditions which render the scope and the accuracy of the interpretation of the results somewhat doubtful. The first is a hypothetical slowing up of the rate of growth somewhere in the adolescent period. The second, which complicates and renders difficult of interpretation the facts in regard to this retardation in growth, grows out of the nature of the tests themselves. It is quite apparent from a number of the test norms that there may be an apparent retardation in rate of growth which is due solely to the organization of the test itself. See for example, the breaks in the curves shown in Figures 6 and 8 (curve 1) which come before the adolescent period and at ages when the growth is shown by other tests to be proceeding at an undiminished rate. It is, therefore, a hazardous proceeding to assume that the rate of growth diminishes at any particular point on the ground of the results from one test alone. Only a still more extensive investigation than has thus far been made will indicate the point at which the rate does diminish and the degree of diminution which comes at this and later ages. The preponderance of evidence, however, seems to indicate that up to some age in early adolescence at least the rate of growth is approximately uniform.

A test which represented in its norms this type of growth, then, could not legitimately use the IQ, unless the other condition were present, namely, a divergence of lines of growth of individuals at different levels or, in other words, an increase in the range of distribution at succeeding ages. We may then examine this matter in the case of those tests for which the distribution at each age is given. The measures of distribution have not been given for all the tests for which norms have been furnished. We have them, however—the Pressey cross-out test (Figure 6), the Otis test (Figure 7), and the Pintner non-language test (Figure 9). The interquartile range has also been calculated for the Chicago Intelligence Test at the various grades. It is as follows:

Grade.....	VII	VIII	IX	X	XI	XII
Range	15.0	11.5	17.6	14.9	16.0	15.9.

In the case of the Illinois Examination, a table is given for the calculation of an individual's IQ which implies a regular and uniform divergence in the scores at succeeding ages. This table is doubtless based upon data from the results of the examination, but

Fig. 4. Age medians in the Yerkes Scale

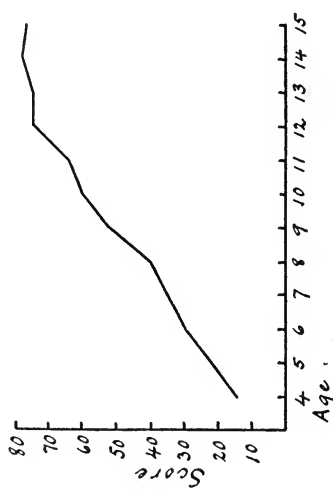


Fig. 5. Age medians for boys in the Pressey Mental Survey Scale No. 1

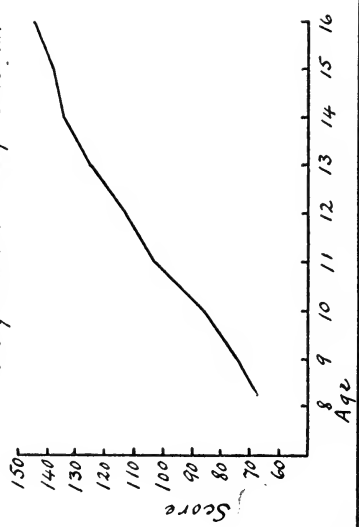


Fig. 6. Curves for the various percentile groups in the Pressey Cross-out Test

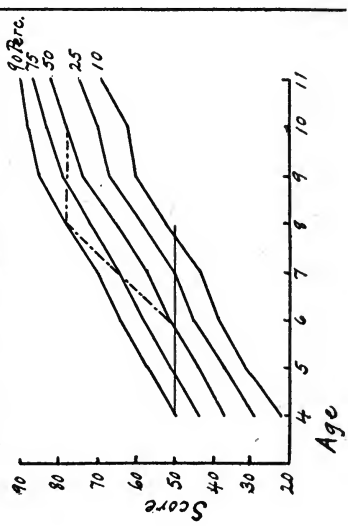
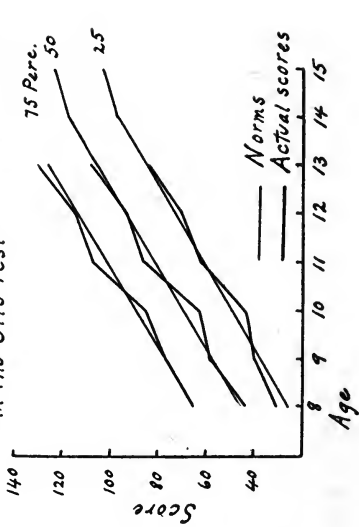


Fig. 7. Curves for three percentile groups in the Otis Test



as these data are not supplied in their original form we shall confine our attention to those cases in which we have the original scores.

Of the four examples which we have before us, which are based on an adequate number of cases, three of them indicate parallel lines of growth, namely, the Pressey test, the Otis test, and the Chicago test. In the Pintner test, the seventy-five percentile curve approaches more closely to the median curve at the higher ages. The twenty-five percentile curve, however, diverges from the median up to age twelve and then remains parallel. It should be noted that the twenty-five percentile curve comes very close to the zero score in the lower ages and this limit in possible scores may be responsible for the fact that it approaches the median at these points. In fact it would be impossible to make a score below year ten which diverges as widely from the median as does the score of the twenty-five percentile at the year twelve, and at the year ten the score would be only slightly over twenty points out of a possible 500 plus. It appears likely therefore that the character of the distribution and possibly the course of the median scores in the early ages is due to the difficulty of the test for the younger children.

The composite effect of the various apparently rather fortuitous factors in the determination of the curves from the Pintner test may be indicated by the following measure of overlapping. We may compare the median score for each age with the percentile scores of the ages next above and below it. If these percentiles approach the fifty percentile (median) at the higher ages the overlapping increases. The facts are shown as follows:

Percentile of age higher or lower which corresponds to the median score at successive ages.

Age	6	7	8	9	10	11	12	13	14
Age higher	40	42	45	45	43	43	46	47	47
Age lower	63	59	57	55	57	57	54	53	53

It may be seen that there is very slight evidence of increasing overlapping from ages 7 to 11. The score for age 6 is too erratic to be relied on and the scores above age 12 follow a different course than do those below.

It may be remarked, parenthetically, that Yerkes' *coefficient of intelligence* implies an increase in the range of distribution with

advancing age, and if this increase does not occur this index is invalid.

It appears from these facts that both of the assumptions which may serve to explain the validity of the IQ in the case of the Binet scale are in question. There seems to be evidence of considerable weight that the typical intellectual growth follows a uniform rate, within at least the period covered roughly by the elementary school. We are here, of course, assuming that the measure of mental ability is a composite one, which gives opportunity for the exercise of a variety of functions, and for a wide range of scores ranging from considerably below the ability of the youngest children tested to considerably above the ability of the oldest. The evidence seems further to indicate that the individuals above and below the median develop at about the same rate, though in different levels as compared with the median child. At any rate, the correctness of the usual assumptions is called in question by the facts before us.

This last tentative conclusion may receive further comment in view of practical experience which seems to contradict it. Our everyday observation seems to indicate that the superior child advances more rapidly than the median child and the inferior child more slowly. In many experiments, for example, it has been shown that the upper fifth, or thereabouts, of a grade may be segregated, and may make much more rapid progress than the mass of the children who remain within the ordinary grade. This, however, may be susceptible of an entirely different explanation than that the superior children are developing with unusual rapidity.

An illustration may be taken from Figure 6. Assume that a group of ninety percentile children, which represents the median of the upper fifth, started in school at the age of six with a group of children of median ability. The rate of advancement in school of these median children might be represented by the dot and dash line. It will be seen in the figure that this line meets the ninety percentile line at eight years, meaning that this group of children has, by the time they are eight years old, and after two years' schooling, reached a degree of school attainment which corresponds to their native ability. This school attainment is seen to be equivalent to that which the median children have reached at ten years, after four years' training.

In other words, these superior children have, as in the experiments alluded to, made four years' advancement in two. This is

not, however, because their mental growth has been twice as rapid as the others, but because in the space of two years' time they have risen to their mental level in school work. A similar illustration might be drawn from the inferior children, who, by retardation and repeating grades, sink to their own mental level.

This is the condition which results from the entrance of children of all mental levels at the same age and from giving them all the same grade of work. If, on the contrary, all children entered school at the same mental level, which is represented by the horizontal line opposite score 50, they would enter at widely different ages, the ninety percentile children entering approximately at four, the seventy-five percentile at five, the median at six, the twenty-five percentile at seven, and the ten percentile at eight. Assuming that this were done and that they were all given the same curriculum, the data before us would indicate that they would advance at equal rates, up to the period at least of the retardation or cessation of mental growth. Assuming that this cessation came at the same age for the different groups, it would come at different levels in school attainment, the ninety percentile child being able to continue his advancement four grades farther than the ten percentile child.

The assumption is not made that this conclusion is strictly accurate. It indicates, however, that it is quite possible to interpret the main facts of our ordinary observation on the basis of the assumption of straight line development, and parallel development at different levels, and this strengthens the conclusion that the assumption that decreasing rate and diverging lines represent the nature of mental growth must be called into serious question.

Our discussion started with the concrete problem of the IQ and with the assumptions which underly it. It has led us to conclude that the application of the IQ to other than the Binet scale must be made with great caution and only after determining that it is a suitable method of representing the scores in other tests. We have been led beyond this practical issue to the consideration of the general nature of mental development and have been led to call in question the time honored conceptions in regard to it. The evidence at hand is sufficient to raise serious doubts about these conceptions and to stimulate to further inquiry to establish on firmer foundation the general theories concerning mental growth.

THE CONSTITUTION OF ARITHMETICAL ABILITIES

EDWARD L. THORNDIKE,
Teachers College, Columbia University.

When the analysis of the mental functions involved in arithmetical learning is made thorough it turns into the question, 'What are the elementary bonds or connections that constitute these functions?'; and when the problem of teaching arithmetic is regarded, as it should be in the light of present psychology, as a problem in the development of a hierarchy of intellectual habits, it becomes in large measure a problem of the choice of the bonds to be formed and of the discovery of the best order in which to form them and the best means of forming each in that order.

The Importance of Habit Formation.—The importance of habit formation or connection-making has been grossly underestimated by the majority of teachers and writers of text-books. For, in the first place, mastery by deductive reasoning of such matters as 'carrying' in addition, 'borrowing' in subtraction, the value of the digits in the partial products in multiplication, the manipulation of the figures in division, the placing of the decimal point after multiplication or division with decimals, or the manipulation of the figures in the multiplication and division of fractions is impossible or extremely unlikely in the case of children of the ages and experience in question. They do not as a rule deduce the method of manipulation from their knowledge of decimal notation. Rather they learn about decimal notation by carrying, borrowing, writing the last figure of each partial product under the multiplier which gives that product, etc., etc. They learn the method of manipulating numbers by seeing them employed, and by more or less blindly acquiring them as associative habits.

In the second place, we, who have already formed and long used the right habits and are thereby protected against the casual misleadings of unfortunate mental connections, can hardly realize the force of mere association. When a child writes sixteen as 61, or finds 428 as the sum of 15, or gives 642 as an answer to

19
16
18
—

27×36 , or says that 4 divided by $\frac{1}{4} = 1$, we are tempted to consider him mentally perverse, forgetting or perhaps never having understood that he goes wrong for exactly the same general reason that we go right, namely, the general law of habit formation. If we study the cases of 61 for 16, we shall find them occurring in the work of pupils who, after having been drilled in writing 26, 36, 46, 62, 63, and so on, in which the order of the six in writing is the same as it is in speech, return to writing the 'teen number. If our language said onety-one for eleven and onety-six for sixteen, we should probably never find such errors except as 'lapses' or as the results of misperception or lack of memory. They would then be more frequent *before* the 20's, 30's, etc., were learned.

If pupils are given much drill on written single column addition involving the higher decades (each time writing the two-figure sum), they are forming a habit of writing 28 after the sum of 8 6 9 and 5 is reached, and it should not surprise us if the pupil still occasionally writes the two figure sum for the first column though a second column is to be added also. On the contrary, unless some counter force influences him, he is absolutely sure to make this mistake.

In connection with the third mistake quoted there is opportunity for a very instructive experiment, namely, to give to a group of children who are just to be taught 'long' multiplication a set of examples where say a three place number is to be multiplied by a three place number, asking them to multiply. Some will do nothing. Some will subtract. Some will multiply them as shown above. Some will multiply the upper number by one or more figures of the lower number. All will exemplify the action of the law of association. They do what they have done in the situation most like the present. Since they have rarely added, never multiplied or divided and have often subtracted with two three place numbers, some will be led by habit to neglect the request to multiply. Since they have multiplied 7 by 6 getting 42, and 3 and 2 getting 6, and 4 by 5 getting 20, some will write 37 and so on. It is with or

26

642

against the force of such habits as these that the new habits of 'long' multiplication are formed.

The last mistake quoted ($4 \div \frac{1}{4} = 1$) is interesting because here we have possibly one of the cases where deduction from psychology alone can give constructive aid to teaching. Multiplication and division by fractions have been notorious for their difficulty. The former is now alleviated by using *of* instead of *X* until the new habit is fixed. The latter is still approached with elaborate caution and with various means of showing why one must 'invert and multiply' or 'multiply by the reciprocal'.

But in the author's opinion it seems clear that the difficulty in multiplying and dividing by a fraction was not that children felt any logical objections to cancelling or inverting. I fancy that the majority of them would cheerfully invert any fraction three times over or cancel numbers at random in a column if they were shown how to do so. But if you are a youngster inexperienced in numerical abstractions and if you have had *divide* connected with 'make smaller' three thousand times and never once connected with 'make bigger', you are sure to be somewhat impelled to make the number smaller the three thousand and first time you are asked to divide it. Some of my readers will probably confess that even now they feel a slight irritation or doubt in saying or writing that $\frac{1}{4} \div \frac{1}{8} = 128$.

The habits that have been confirmed by every multiplication and division by integers are, in this particular of "*the ratio of result to number operated upon*", directly opposed to the formation of the habits required with fractions. And that is, I believe, the main cause of the difficulty. Its treatment then becomes easy, as will be shown later.

These illustrations could be added to almost indefinitely, especially in the case of the responses made to the so-called 'catch' problems. The fact is that the learner rarely can, and almost never does, survey and analyze an arithmetical situation and justify what he is going to do by articulate deductions from principles. He usually feels the situation more or less vaguely and responds to it as he has responded to it or some situation like it in the past. Arithmetic is to him not a logical doctrine which he applies to various special instances, but a set of rather specialized habits of behavior toward certain sorts of quantities and relations. And in so far as he does come to know the doctrine it is chiefly by doing the will of the master. This is true even with the clearest exposi-

tions, the wisest use of objective aids and full encouragement of originality on the pupil's part.

Lest the last few paragraphs be misunderstood, I hasten to add that the psychologists of today do not wish to make the learning of arithmetic a mere matter of acquiring thousands of disconnected habits, nor to decrease by one jot the pupil's genuine comprehension of its general truths. They wish him to reason not less than he has in the past, but more. They find, however, that you do not secure reasoning in a pupil by demanding it, and that his learning of a general truth without the proper development of organized habits back of it is likely to be, not a rational learning of that general truth, but only a mechanical memorizing of a verbal statement of it. They have come to know that reasoning is not a magic force working in independence of ordinary habits of thought, but an organization and coöperation of those very habits on a higher level.

The older pedagogy of arithmetic stated a general law or truth or principle, ordered the pupil to learn it, and gave him tasks to do which he could not do profitably unless he understood the principle. It left him to build up himself the particular habits needed to give him understanding and mastery of the principle. The newer pedagogy is careful to help him build up these connections or bonds ahead of and along with the general truth or principle, so that he can understand it better. The older pedagogy commanded the pupil to reason and let him suffer the penalty of small profit from the work if he did not. The newer provides instructive experiences with numbers which will stimulate the pupil to reason so far as he has the capacity, but will still be profitable to him in concrete knowledge and skill, even if he lacks the ability to develop the experiences into a general understanding of the principles of numbers. The newer pedagogy secures more reasoning in reality by not pretending to secure so much.

The newer pedagogy of arithmetic, then, scrutinizes every element of knowledge, every connection made in the mind of the learner, so as to choose those which provide the most instructive experiences, those which will grow together into an orderly, rational system of thinking about numbers and quantitative facts. It is not enough for a problem to be a test of understanding of a principle; it must also be helpful in and of itself. It is not enough for an

example to be a case of some rule; it must help review and consolidate habits already acquired or lead up to and facilitate habits to be acquired. Every detail of the pupil's work must do the maximum service in arithmetical learning.

It is then profitable to consider detailed cases of bonds now often neglected which deserve attention. Those listed below are among the most important.

Numbers As Measures of Continuous Quantities.—The numbers one, two, three, 1, 2, 3, etc., should be connected soon after the beginning of arithmetic each with the appropriate amount of some continuous quantity like length or volume or weight, as well as with the appropriate sized collection of apples, counters, blocks and the like. Lines should be labelled 1 foot, 2 feet, 3 feet, etc.; one inch, two inches, three inches, etc.; weight should be lifted and called one pound, two pounds, etc.; things should be measured in glassfuls, handfuls, pints, and quarts. Otherwise the pupil is likely to limit the meaning of, say, four to four sensibly discrete things and to have difficulty in multiplication and division. Measuring or counting by insensibly marked off repetitions of a unit, binds each number name to its meaning as—times *whatever* 1 is, more surely than mere counting of the units in a collection can, and should reinforce the latter.

(2) *Additions in the Higher Decades.*—In the case of all save the very gifted children, the additions with higher decades—that is, the bonds, $16 + 7 = 23$, $26 + 7 = 33$, $36 + 7 = 43$, $14 + 8 = 22$, $24 + 8 = 32$, and the like—need to be specifically practiced until the tendency becomes generalized. 'Counting' by 2s beginning with 1, and with 2, counting by 3s beginning with 1, with 2, and with 3, counting by 4s beginning with 1, with 2, with 3, and with 4, and so on, make easy beginnings in the formation of the decade connections. Practice with isolated bonds should soon be added to get freer use of the bonds. The work of column addition should be checked for accuracy so that a pupil will continually get beneficial practice rather than 'practice in error'.

(3) *The Uneven Divisions.*—The quotients with the remainder for the divisions of every number to 19 by 2, every number to 29 by 3, every number to 39 by 4, and so on should be taught as well as the even divisions. A table like the following will be found a convenient means of making these connections:

$$\begin{aligned}
 10 &= \dots 2s. \\
 10 &= \dots 3s \text{ and } \dots \text{rem.} \\
 19 &= \dots 4s \text{ and } \dots \text{rem.} \\
 10 &= \dots 5s. \\
 11 &= \dots 2s \text{ and } \dots \text{rem.} \\
 11 &= \dots 3s \text{ and } \dots \text{rem.} \\
 &\cdot \\
 &\cdot \\
 &\cdot \\
 89 &= \dots 9s \text{ and } \dots \text{rem.}
 \end{aligned}$$

These bonds must be formed before short division can be efficient, are useful as a partial help toward selection of the proper quotient figures in long division, and are the chief instruments for one of the important problem series in applied arithmetic,—“How many X’s can I buy for y cents at z cents per x and how much will I have left?” That these bonds are at present sadly neglected is shown by Kirby ('13) who found that pupils in the last half of grade 3 and the first half of grade 4 could do only about four such examples per minute (in a ten-minute test), and even at that rate made far from perfect records, though they had been taught the regular division tables. Sixty minutes of practice resulted in a gain of nearly 75 per cent in number done per minute, with an increase in accuracy as well.

(4) *The Equation Form.*—The equation form with an unknown quantity to be determined, or a missing number to be found should be connected with its meaning and with the problem attitude long before a pupil begins algebra, and in the minds of pupils who never will study algebra.

Children who have just barely learned to add and subtract learn easily to do such work as the following:

$$\begin{aligned}
 4 + 8 &= \dots \\
 5 + \dots &= 4. \\
 \dots + 3 &= 11. \\
 \dots &= 5 + 2. \\
 16 &= 7 + \dots \\
 12 &= \dots + 5.
 \end{aligned}$$

The equation form is the simplest uniform yet devised to state a quantitative issue. It is capable of indefinite extension if certain easily understood conventions about parenthesis and fraction signs are learned. It should be employed widely in accounting and the treatment of commercial problems, and would be except for outworn conventions. It is the chief contribution of algebra to business and industrial life and one that arithmetic can make as well. It saves more time in the case of drills on reducing fractions to higher and lower terms alone than was required to learn its meaning and use. To rewrite a quantitative problem as an equation and then make the easy selection of the necessary technique to solve the equation is one of the most universally useful intellectual devices known to man. The words 'equals', 'equal', 'is', 'are', 'makes', 'make', 'gives', 'give', and their rarer equivalents should therefore early give way on many occasions to the '=' which so far surpasses them in ultimate convenience and simplicity.

(5) *Addition and Subtraction Facts in the Case of Fractions.*—

In the case of adding and subtracting fractions, certain specific bonds—between the situation of halves and thirds to be added and the responses of thinking of the numbers as equal to so many sixths, between the situation thirds and fourths to be added and thinking of them as so many twelfths, between fourths and eighths to be added and thinking of them as eighths, and the like—should be formed separately. The general rule of thinking of fractions as their equivalent with some convenient denominator should come as an organization and extension of such special habits, not as an edict from the text-book or teacher.

(6) *Fractional Equivalents.*—Efficiency requires that in the end the much used reductions should be firmly connected with the situations where they are needed. They may as well, therefore, be so connected from the beginning with the gain of making the general process far easier for the dull pupils to master. We shall see later that, for all save the very gifted pupils, the economical way to get an understanding of arithmetical principles is not, usually, to learn a rule and then apply it, but to perform instructive operations and, in the course of performing them, to get insight into the principles.

(7) *Protective Habits in Multiplying and Dividing with Fractions.*—In multiplying and dividing with fractions special bonds should be formed to counteract the now harmful influence of the

'multiply = get a larger number' and 'divide = get a smaller number' bonds which all work with integers has been reinforcing.

For example, at the beginning of the systematic work with multiplication by a fraction, let the following be printed clearly at the top of every relevant page of the text-book and displayed on the blackboard.

When you multiply a number by anything more than 1 the result is larger than the number.

When you multiply a number by 1 the result is the same as the number.

When you multiply a number by anything less than 1 the result is smaller than the number.

Let the pupils establish the new habit by many such exercises as:—

$18 \times 4 = \dots$	$9 \times 2 = \dots$
$4 \times 4 = \dots$	$6 \times 2 = \dots$
$2 \times 4 = \dots$	$3 \times 2 = \dots$
$1 \times 4 = \dots$	$1 \times 2 = \dots$
$\frac{1}{2} \times 4 = \dots$	$\frac{1}{3} \times 2 = \dots$
$\frac{1}{4} \times 4 = \dots$	$\frac{1}{6} \times 2 = \dots$
$\frac{1}{8} \times 4 = \dots$	$\frac{1}{6} \times 2 = \dots$

In the case of division by a fraction the old harmful habit should be counteracted and refined by similar rules and exercises as follows:

When you divide a number by anything more than 1 the result is smaller

When you divide a number by 1 the result is the same as the number.

When you divide a number by anything less than 1 the result is larger than the number.

State the missing numbers:—

$8 = \dots 4s$	$12 = \dots 6s$	$9 = \dots 9s$
$8 = \dots 2s$	$12 = \dots 4s$	$9 = \dots 3s$
$8 = \dots 1s$	$12 = \dots 3s$	$9 = \dots 1s$
$8 = \dots \frac{1}{2}s$	$12 = \dots 2s$	$9 = \dots \frac{1}{3}s$
$8 = \dots \frac{1}{4}s$	$12 = \dots 1s$	$9 = \dots \frac{1}{6}s$
$8 = \dots \frac{1}{8}s$	$12 = \dots \frac{1}{2}s$	
	$12 = \dots \frac{1}{3}s$	
	$12 = \dots \frac{1}{4}s$	

$16 \div 16 =$	$9 \div 9 =$	$10 \div 10 =$	$12 \div 6 =$
$16 \div 8 =$	$9 \div 3 =$	$10 \div 5 =$	$12 \div 4 =$
$16 \div 4 =$	$9 \div 1 =$	$10 \div 1 =$	$12 \div 3 =$
$16 \div 2 =$	$9 \div \frac{1}{3} =$	$10 \div \frac{1}{5} =$	$12 \div 2 =$
$16 \div 1 =$	$9 \div \frac{1}{9} =$	$10 \div \frac{1}{10} =$	$12 \div 1 =$
$16 \div \frac{1}{2} =$			$12 \div \frac{1}{2} =$
$16 \div \frac{1}{4} =$			$12 \div \frac{1}{3} =$
$16 \div \frac{1}{8} =$			$12 \div \frac{1}{4} =$
			$12 \div \frac{1}{6} =$

(8) *Per Cent of Means 'Hundredths Times'.*—In the case of percentage a series of bonds like the following should be formed:

5 per cent of	=	.05 times
20 " " "	=	.20 "
6 " " "	=	.06 "
25%	=	.25 ×
12%	=	.12 ×
3%	=	.03 ×

Four five-minute drills on such connections between 'X per cent of' and 'its decimal equivalent times' are worth an hour's study of verbal definitions of the meaning of per cent as per hundred or the like. The only use of the study of such definitions is to facilitate the later formation of the bonds, and, with all save the brighter pupils, the bonds are more needed for an understanding of the definitions than the definitions are needed for the formation of the bonds.

(9) *Habits of Verifying Results.*—Bonds should early be formed between certain manipulations of numbers and certain means of checking, or verifying the correctness of, the manipulation in question. The additions to $9 + 9$ and the subtractions to $18 - 9$ should be verified by objective addition and subtraction and counting until the pupil has sure command; the multiplications to 9×9 should be verified by objective multiplication and counting of the result (in piles of tens and a pile of ones) eight or ten times,* and by addition eight or ten times;* the divisions to $81 \div 9$ should be verified by multiplication and occasionally object-

*Eight or ten times *in all*, not eight or ten times for each fact of the tables.

ively until the pupil has sure command: column addition should be checked by adding the columns separately and adding the sums so obtained and by making two shorter tasks of the given task and adding the two sums; 'short' multiplication should be verified eight or ten times by addition; 'long' multiplication should be checked by reversing multiplier and multiplicand and in other ways; 'short' and 'long' division should be verified by multiplication.

These habits of testing an obtained result are of threefold value. They enable the pupil to find his own errors, and to maintain a standard of accuracy by himself. They give him a sense of the relations of the processes and the reasons why the right ways of adding, subtracting, multiplying and dividing are right, such as only the very bright pupils can get from verbal explanations. They put his acquisition of a certain power, say multiplication, to a real and intelligible use, in checking the results of his practice of a new power, and so instill a respect for arithmetical power and skill in general. The time spent in such verification produces these results at little cost; for the practice in adding to verify multiplications, in multiplying to verify divisions, and the like is nearly as good for general drill and review of the addition and multiplication themselves as practice devised for that special purpose.

Early work in adding, subtracting, and reducing fractions should be verified by objective aids in the shape of lines and areas divided in suitable fractional parts. Early work with decimal fractions should be verified by the use of the equivalent common fractions for .25, .75, .125, .375, and the like. Multiplication and division with fractions both common and decimal should in the early stages be verified by objective aids. The placing of the decimal point in multiplication and division with decimal fractions should be verified by such exercises as:—

20

1.23 $\overline{) 24.60}$. It cannot be 200; for 200×1.23 is much more than 246.

24.6. It cannot be 2; for 2×1.23 is much less than 24.6.

The establishment of habits of verifying results and their use is very greatly needed. The percentage of wrong answers in arithmetical work in schools is now so high that the pupils are often being practiced in error. In many cases they can feel no genuine

and effective confidence in the processes, since their own use of the processes brings wrong answers as often as right. In solving problems they often cannot decide whether they have done the right thing or the wrong, since even if they have done the right thing, they may have done it inaccurately. A wrong answer to a problem is therefore far often ambiguous and uninformative to them.

These illustrations of the last few pages are samples of the procedures recommended by a consideration of all the bonds that one might form and of the contribution that each would make toward the abilities that the study of arithmetic should develop and improve. It is by doing more or less at haphazard what a psychology teaches us to do deliberately and systematically in this respect that many of the past advances in the teaching of arithmetic have been made.

A CRITICAL STUDY OF THE CONCEPT OF SILENT READING ABILITY

L. W. PRESSEY and S. L. PRESSEY.

University of Indiana.

I. THE MISLEADING CHARACTER OF MANY TEST "LABELS"

As has been well pointed out recently¹ by Thorndike and Courtis, the name which is applied to a test may often be positively misleading, in its implications as to what functions that test measures. A test consisting of problems in addition might seem unmistakably a test of ability to add; but careful investigation may discover that comparative standing on this test is determined almost wholly by speed in writing the answers, or knack in working ahead of one's pencil. And the determination of true ability in addition from any addition test is found to be by no means a simple matter.

If such difficulties appear in the case of such a very concrete and specific subject as arithmetic it is surely reasonable to infer even greater likelihood of error in efforts at measurement in such subjects as literature, composition, silent reading, and so on. Particularly (the writers have felt) was the situation confused and ill-defined with regard to tests in silent reading, and to "silent reading ability." There are here no very specific habits, as in measurement of ability in the fundamentals of arithmetic, which make very specific materials for investigation. Nor are there—as "silent reading ability" is usually conceived—any informational elements, as in history or geography, about which investigation may be centered. Under such circumstances, the formulation of tests which would in truth measure "silent reading ability" would seem a peculiarly difficult task. It is surely easy enough to devise a test in which the element superficially most prominent is ability to read with understanding. We may then, with a certain nonchalance, label such a test a "test of silent reading ability." But to a much greater extent than with addition tests is there danger that our "label" may be inaccurate and even harmfully misleading. Critical research with regard to such tests would seem, then, all the more needful.

¹Thorndike E. L. and Courtis S. A., "Correction Formulae for Addition Tests," *Teachers' College Record*, 1920, 21, 1-24.

It is such a critical study that the present paper reports. Previous research with silent reading tests had led the writers to feel that the situation, with regard to silent reading, needed to be analyzed. Particularly, they felt that ability in assimilative reading was conditioned much more by the nature of the material read than is ordinarily supposed. The problem of the present paper may, then, be put as follows: *is either the form or the content of the matter read an important conditioning factor in silent reading?* If the nature of the material read should appear to be a dominant feature in the situation, it would seem that our tests of general silent reading ability were badly mislabeled.

II. MATERIALS AND RESULTS

The materials of this study were obtained with four reading "scales" each composed of a definite type of reading material. The items for three of them were taken from the Monroe Reading Scales and the "Illinois Examination"²; some of the items in the fourth "scale" also came from these two scales. Each "scale" had twelve items. One "scale" was made up entirely of poetry, one of scientific passages and two of ordinary reading matter, stories, and the like. The two "general" scales were made by simply taking alternate items from the Illinois Examination for the 3rd, 4th and 5th grades, the same examination for the 6th, 7th and 8th grades and from the two forms of Monroe's reading tests for the 6th, 7th and 8th grades. The items involving poetry were omitted. These two scales were roughly equal in difficulty, and of a very similar type of reading matter. An equivalent "scale" was then made up using the poetical items. Very few "scientific" items could be found in the scales, but those that appear were used, others being made to fill out the necessary number of items for a fourth "scale". The distribution of scores and the medians for these four scales were practically the same; the medians of the four are,—6.1, 6.5, 6.6, 6.4. It seems reasonable to suppose, then, that these scales are of about equal difficulty. Samples from these four "scales" appear below:

From the	O suns and skies and clouds of June,
poetical	And flowers of June together,
scale	You cannot rival for one hour
	October's bright blue weather.

²See Journal of Educational Research, October, 1920. The Illinois Examination contains a revision of the original Monroe Scale.

Which month does this stanza say is the more pleasant?

April September May June October

From the
first
"general"
scale

The caravan, stretched out upon the desert, was very picturesque; in motion, however, it was like a lazy serpent. By and by its stubborn dragging became intolerably irksome to Balthasar, patient as he was.

Place a line under the word which tells in what respect the caravan resembled a serpent.

temper color length motion size

From the
second
"general"
scale

It was the garden-land of Antioch. Even the hedges, besides the lure of the shade, offered passers-by sweet promises of wine and clusters of purple grapes. Over melon patches and through apricot and fig groves and groves of oranges and lime, the white-washed houses of the farmers were seen.

What kind of a land was this? Draw a line under the correct answer.

barren hilly productive infertile desert

From the
"scientific"
scale

The tighter a wire is stretched, the higher the tone produced will be when the wire is struck. Five wires are stretched with weights on the ends of them. One weight is 100 lbs., one is 75 lbs., one is 25 lbs., one is 20 lbs. and one is 15 lbs.

Underline the number of lbs. in the weight that will cause the highest tone.

75 lbs. 100 lbs. 25 lbs. 20 lbs. 15 lbs.

These four "scales" were then given to a large seventh grade,^a giving a total of 112 cases. The "poetical" scale was given first, followed by the first "general" scale; then came the "scientific"

^aThe writers wish to express their obligations to Superintendent E. W. Montgomery of Bedford, Indiana, for his kindness and co-operation in this research.

scale and the second "general" scale. By this arrangement, practice with the tests, fatigue and so on were so distributed that they could not act as a constant error in the results. The tests were scored simply on the basis of the total number of items right with each type of reading matter, previous research having suggested that "rate" score was negligible as a measure of silent reading ability, independent of comprehension score.⁴ All possible correlations between the four "scales" were then found. The coefficients appear below:⁵

Correlation of "general 1" with "general 2".85
Correlation of "general 1" and "poetry".38
Correlation of "general 1" and "scientific".35
Correlation of "general 2" and "poetry".31
Correlation of "general 2" and "scientific".49
Correlation of "scientific" and "poetry".56

The exceedingly high correlation between the two "general" scales is first to be noted. This coefficient gives a measure of the "reliability" or "consistency" of the scores yielded by the scale. It seems reasonable to believe that two "poetry" or two "scientific" scales would yield scores that correlated to somewhat the same extent. Similar types of reading matter would probably correlate highly. The writers happen to know that this coefficient of reliability is somewhat higher than is usually found with the Monroe scales,—the coefficients usually run from .65 to .75. It might be suggested that it is the presence of items involving poetry or scientific reading that lowers these correlations.

The striking fact, however, is the contrast between the very high correlation above noted and the remaining coefficients. Strictly, the most reliable coefficients are those between "general 1" and the "scientific", and between "general 1" and "poetry", as contrasted

⁴Pressey, S. L. and L. W., "The Relative Value of Rate and Comprehension Scores on Monroe's Silent Reading Test." *School and Society*, June, 1920. Since that paper was published, a second study has been made confirming the finding as to the negligible nature of rate score. *It also indicates that Monroe's weighting of his comprehension scores adds practically nothing to a crude score as a measure of reading ability.* As a result of these two studies, crude scores,—that is, comprehension scores—have been used throughout this study with no attempt at any more elaborate treatment.

⁵It should be added, in this connection, that a preliminary trial of the four scales in a consolidated country school, testing grades 5, 6, 7 and 8, gave practically identical results. Only 67 children were included in the correlation, and the group was heterogeneous, so the correlations are not as reliable as those given in the paper. But they make very pretty confirmatory evidence. The correlation between the two "general" scales was .83, and none of the others were above .49.

with the correlations between "general 1" and "general 2"; the most significant contrast appears here. That is, the "poetry" scale immediately preceded "general 1" while the "scientific" passages immediately followed it. Adventitious circumstances such as fatigue should, then, influence these correlations less than those between scales further apart in the series. The correlation between "general 1" and "general 2" is, thus, a correlation between scales somewhat apart in the series; this correlation is thus probably slightly lower than it should be, while the correlations between "general 1" and the other two scales express to the full the relation between ability to read differing types of subject matter.

The results are surely striking. The findings need confirmation of course. Study of reading ability in its relation to vocabulary in a particular field, study of ability in assimilative reading interesting and not interesting to a given child, studies in "assimilative listening" as compared with ability in assimilative reading,—there are a dozen and one bits of investigation which at once suggest themselves as likely to throw further light on the problem. Meanwhile, certain implications of the data just presented are worth considering.

III. NEED FOR A RE-FORMULATION OF THE SILENT READING PROBLEM

It appears, then, that ability in silent reading depends very largely upon the nature of the passage read; a good reader in one type of subject matter may very likely be a poor reader with other material. And the usual silent reading scale may be considered to measure—certainly not silent reading ability in general, since there seems to be little evidence of any general factor of outstanding importance. Nor do the standard scales specialize on one type or other of subject matter; it might possibly be questioned whether they measured much more than ability in dealing with such catechistic series of questions. The writers would hardly go to such an extreme conclusion. But they do feel that the results suggest a distinct revision of present concepts with regard to silent reading ability and its measurement. As the writers see it, "silent reading ability" involves four distinct factors.

There must be (1) freedom from oral reading habits. It is evidently of great importance to achieve such freedom. This factor is a negative one, however, and is not to be considered correlative

with a positive ability. Tests for detecting the presence of oral reading habits would hardly take the form of the usual reading test. Further, after freedom from such handicaps has been won, the factor should be negligible; one is then above the "threshold" in silent reading ability. Silent reading ability, defined as assimilative reading independent of oral reading habits, needs no more to be stressed than phonetics in the sixth grade.

A second factor (2) in efficient assimilative reading is a large reading vocabulary and a background of information. Evidently a reading vocabulary is the factor most readily acquired by reading. Measurement of vocabulary should, the writers feel, be a regular practice in appraising the work of a teacher or school. By a curious and unfortunate circumstance, vocabulary tests have received most attention as measures of general intelligence. The acquirement of an adequate reading vocabulary, presumably by a large amount of required reading of a type in which the pupils are interested, would seem to the writers a matter to which the schools should give decided attention. In fact, they feel that this is the neglected element in the silent reading situation.⁶

A third factor (3) of prime importance is, of course, the development of interest to motivate reading of the right kind. Once there is such interest, a pupil will usually shake himself free of oral reading habits in his eagerness to cover the subject matter in which he is absorbed. And if there is interest, the acquirement of a reading vocabulary will come of itself. The writers have a distinct feeling that silent reading is now a problem in the public schools largely because the reading matter required by the schools is not such as to interest the pupils. If there had been interest, the two previously mentioned factors would, in most instances, take care of themselves.

The fourth factor (4) which conditions efficiency in assimilative reading may be considered the development of habits of attention and application. But these factors are factors in silent reading only incidentally and training for this factor in silent reading is, of course, training in methods of study.

What, then, about the present tests in silent reading? It is the

⁶It might be added that the writers have included such a test in a scale, designed to measure achievement in the eighth grade. A full statement with regard to this scale will be published shortly. The material reported in the present paper was, in fact, gathered in connection with the construction of this scale.

writers' guess that scales of the type of the Kansas Test and the Monroe Test are really by far the best examples so far of tests of attention,—which the devisors have stumbled upon without knowing it. They are good tests, but they have the wrong label. They are useful in investigating the efficiency of study habits. They may be used in hunting out those who are not yet free from oral reading habits—though instruments more explicitly devised for this one purpose should be much more efficient. For investigation of real ability in assimilative reading the writers would suggest: (a) a preliminary instrument for detecting oral reading habits; and (b) a test of vocabulary,—a classified test covering specific fields of information would be most valuable. These two tests might well be supplemented by (c) any one of the present standard reading tests. It should be clearly understood, however, that these last tests are used primarily to investigate habits of attention, not any process peculiar to the reading process alone.

These are the writers' conclusions on the basis of their data. Their findings need confirming. Meanwhile, tentative though these conclusions are, they have an interesting bearing on the general problem. They also suggest certain obvious shifts in present methods of teaching reading, and in interpreting the results with reading tests which are of no little interest.

A COMBINED MENTAL-EDUCATIONAL SURVEY *

RUDOLF PINTNER and HELEN MARSHALL,
Ohio State University.

During the last twenty years there have been two relatively independent movements within the field of psychology which have had to do with measurement, namely, the measurement of mental ability and the measurement of educational attainment. Although overlapping in many respects, it is only recently that any attempt has been made to combine the methods and results of these two lines of investigation in order to answer the question as to how much educational attainment we have a right to expect from a child of a given mental ability.

Children vary greatly in mental ability and our mental tests give us a fairly accurate measure of their native ability. These tests are measures of what children are capable of achieving rather than of what they have achieved. They are measures of the degree of modifiability of the individual. They are measures of the raw material with which the school and the teacher have to work.

Educational scales give us a measure of the amount of attainment in any specific school subject. They measure the extent to which the school has been successful in modifying its pupils. They give us the actual facts as to how much has been attained by each pupil.

These two measures, as they stand, are both extremely valuable. They increase tremendously in significance, however, as soon as we use the two measures in combination, and evaluate the one in the light of the other. If we make such an evaluation, we can expect answers to questions of this kind: How much educational attainment can I expect of this child who possesses so much general intelligence? If a child is able to do so much on our intelligence tests, how much of the ordinary school subjects ought he to know if given the common school instruction? Is this child of high intelligence accomplishing school work commensurate with his mental ability? Or is he working at a lower level, because he is in a grade that does not call for as much as he really could accomplish? Is this

*The two test blanks described in this article, namely, the Non-Language Mental Test and the Educational Survey Test, together with stencils for scoring and Manual of Directions giving tables of norms, may all be obtained from the College Book Store, Columbus, O.

school doing work worthy of the native ability of its pupil-material? Here is a school doing good work as judged by the average school in the system and also by the norms of standard educational tests, but is it really working up to capacity, or is it able to achieve such results easily and without effort because of the superior intelligence of its pupil-material?

We can see from the type of question here suggested, that we are raising fundamental questions, which the intelligence test alone or the educational test alone is unable to answer. We are attempting to use every bit of native ability of the pupil and by so doing, eliminate the waste of intelligence that is going on in our schools today. This wasted intelligence is much greater than most of us have realized. Ever since the introduction of mental and educational tests, we have all been aware of this wastage, but we have not been able to measure it accurately. The right combination of mental and educational tests will enable us to do so. It will show us just where this wastage exists, and its amount. It will show us just where—upon what school or class or pupil—pressure must be brought to bear, in order to eliminate the waste and make the school, class or pupil work up to the extent of the native capacity.

As a matter of fact we shall see, what we have long suspected and what we have only recently consciously realized, that the greatest amount of waste exists among the brighter pupils in a class or among the better schools in a school system. It is, as a rule, the more intelligent pupils that are working below capacity, even although they are keeping well up to the average level of the group. We have been pushing and cramming the duller children, while the brighter ones have been allowed to loaf. The bright child is the most retarded child in our schools. The dull child is the most accelerated. The bright child is the laziest child, and the dull child is the most industrious. We are now ready by means of combined mental and educational tests to equalize the pressure, and stimulate equally both the bright and the dull so that they may work up to their respective capacities. If we can accomplish this we shall have happier and better children. Lack of adequate stimulus undoubtedly leads to much disciplinary trouble in our schools. Habits of mental laziness acquired in school often persist through life, and there are undoubtedly many adults at the present time, who have splendid

native ability and who do not know it, because the school has taught them to be satisfied with a mediocre type of accomplishment.

The Tests.—The use of mental and educational tests in combination for the purposes outlined above was suggested by one of the writers* in 1918, arising naturally out of the plea for a wider use of group tests for survey purposes. Attempts to use the standard educational tests, as they existed at that time, for this combined mental-educational measure were found to be impossible, owing to the fact that practically no age norms for the tests had been published. Being tests of school achievement most authors had been content with grade norms only. In addition to this, most of the educational tests required a long time to give and several hours would be necessary in order to cover the chief subjects of the elementary school curriculum. It was, therefore, decided to devise a short educational survey test covering the chief elementary school subjects. This was composed of eight tests made up from standard educational tests and requiring about 30 minutes for its application. The tests, out of which this short composite educational test was composed, are (1) Thorndike's Vocabulary, (2) Woody's Arithmetic, (3) Kelly's Reading, (4) Thorndike's Reading, (5) Trabue's Language Completion, (6) Starch's Punctuation and Grammar, (7) Hahn and Lackey's Geography, (8) Van Waganen's History. The composition of this test and its relation to the longer standard tests have been discussed elsewhere.**

For a group mental test to go along with this educational survey test, it was decided to construct a test which would have as wide an application to as many different types of children as possible and which would also be far removed from ordinary school work. For this purpose a non-language group test was devised. It has been described in detail elsewhere.*** This test has been found useful for all school children from grade II upwards. It can be given with equal fairness to English-speaking as well as to non-English speaking or deaf children. It does not overlap at all with the type of thing that is taught in school. By avoiding reading and language tests, we can gauge more accurately the intelligence of children who have been deprived of schooling or whose schooling has been

*Pintner, R. *The Mental Survey*, Appletons, 1918.

**Pintner, R., and Fitzgerald, F. An Educational Survey Test. *Journal of Educational Psychology*, Vol. XI. No. 4, April, 1920, pp. 207-223.

***Pintner, R. A. Non-Language Group Intelligence Test. *J. of Applied Psych.* Vol. III. Sept., 1919. pp. 199-214.

very irregular or who have special disabilities in reading and, when we have done so, we can estimate what they ought to be able to accomplish in school. We shall likewise avoid the error of over-rating the intelligence of the bookish child, who has had undue stimulation at home from the reading and language side.

The Standardization.—Standards for such a combined measure must be based upon children who have taken both tests. This is absolutely necessary, because of the wide variability in mentality and in educational accomplishment found among different groups of children. Only when we are sure that our standards are based upon a large and unselected sampling of school children, can we evaluate educational tests in terms of mental tests, and it is very seldom that we can be sure of any such thing. It was, therefore, decided to make a standardization based solely upon those cases that had been given both tests.

Altogether 4303 cases were tested on both tests, ranging in age from 6 to adult. There were from 300 to 600 cases at the intermediate ages and fewer at the upper and lower ends of the standardization. The distribution of the scores on both tests seemed to indicate an approach to a normal distribution when examined age by age. There was likewise a reasonable increase in both tests from one age to the other.

The Method of Rating.—The percentile rating is undoubtedly the most accurate and useful rating for mental testing purposes. In this case, however, we are dealing with two tests and we are primarily concerned with the difference between the ratings in these two tests, in order that we may be able to state by how much a child in his educational accomplishment falls below his mental ability or vice versa. If we have his percentile ratings on the mental and on the educational tests, the difference between these two percentile ratings will be the type of measure we want, except that it will be very inaccurate. In as much as our distribution approaches the normal, a difference of five points near the median will be of much less significance than the same difference at either end of the distribution. For example, if one child makes percentiles of 50 and 45 on the mental and educational tests, respectively, and another 85 and 80, both will show a difference of 5 points. Both are 5 points lower educationally than they are mentally. But the significance of this

difference of 5 points is very different in the two cases. In the center of our distribution there is little difference in actual ability between the 45th and the 50th case; but there is a great difference in ability between the 80th and the 85th case. It is obvious, therefore, that the percentile method, in spite of its many advantages, cannot be used for our purpose.

The method adopted was, therefore, to assume a normal distribution of the scores and convert the percentage frequency of any score into terms of the mean square deviation or sigma. By this means all scores are evaluated in terms of sigma and the difference between these sigma values is comparable at any part of the surface of distribution. This procedure was carried out for both tests for all ages. It would, however, be rather inconvenient and decidedly puzzling to the average teacher to be asked to think in terms of plus and minus sigma, and particularly so when we are handling the results of two tests expressed in sigma values and when we require the difference between these two sigma values to stand for the excess or deficiency in the mental test over the educational.

To avoid this complexity the sigma values were turned into a scale of so-called index numbers running from 0 to 100, zero being placed at the lower end of the distribution and 100 at the other, and 50 corresponding to the zero point of the normal probability surface. If now we multiply the sigma values by a constant and add or subtract the product from 50 we obtain a new scale of numbers, which we have called index values. These values are all positive whole numbers and the teacher can readily understand their significance. It is easy for her to grasp the meaning of a mental index of 50 and an educational index of 47 and a difference between the two of minus three, which can be interpreted as meaning that the particular child is three points lower in educational attainment than he should be as compared with his native ability. And furthermore that this child with a difference of -3 is not nearly so poorly adjusted as another child who shows a difference of -12 after subtracting his educational index of 70 from his mental index of 82, even although the mental ability and educational accomplishment of the latter child are far above those of the first case. In other words, the teacher, after scoring the tests, converts the scores by means of a table into index values and the significance of the index values is explained as follows:

Index Values	Mental	Educational	Per cent of cases
0 to 19	Dull	Very Poor	2
20 to 39	Backward	Poor	23
40 to 59	Normal	Average	50
60 to 79	Bright	Good	23
80 to 100	Very Bright	Very Good	2

That is to say, all index values between 0 and 19 on the mental test show dull mentality and on the educational test indicate very poor educational accomplishment, and fall among the lowest 2 per cent of the cases tested; similarly index values from 20 to 39 indicate backward mentality or poor educational accomplishment and fall among the lower 23 per cent of the cases, and so on. The examiner soon becomes familiar with this table, because it is very easy to remember. The first 20 index numbers refer to the lowest group, the next 20 to the lower group, the next 20 to the average or intermediate group, the next 20 to the higher group and the top 20 to the highest group.

The Three Ratings.—Having given the two tests and having calculated the indices and the difference, the examiner now has three valuable ratings:

- 1) The Mental Index, which is a measure of the native ability of the child.
- 2) The Educational Index, which is a measure of the school attainment of the child.
- 3) The Difference, which expresses the relationship between the two previous ratings.

The first two ratings have already been interpreted and the meaning of an educational or mental rating is now thoroughly familiar to all teachers. The new rating that we have introduced here is what we have called the Difference, i. e., the difference of a child's standing in the educational and mental tests, or the difference between his native capacity and his actual accomplishment. Such a rating we believe to be the most important contribution that the science of mental measurement can offer to the educator. This Difference is arrived at by subtracting the mental from the educational index. If the mental index is higher the Difference is marked by a minus.

If the educational index is higher the Difference is marked by a plus. A minus difference means that the child is doing less educational work than he has the ability to accomplish. A plus difference means that he is doing more educationally than has usually been accomplished by children of like mentality. Notice the significance of the plus difference. Because our educational test norms are based upon what is actually now being accomplished by a fair sampling of school children, we have no measure of what ought to be accomplished under ideal conditions where each child is working up to the limit of his capacity. It is useless to attempt to set up any such ideal standard. We, therefore, find many cases with a plus difference and this means that these cases are children who are accomplishing in school work more than is usually accomplished by children of like intelligence.

The Significance of the Difference Ratings.—In order to interpret these Differences, a frequency distribution of the Differences of 4303 cases was made. We may conveniently divide this frequency distribution into a five-fold grouping, as follows:

Differences	Percentage of Cases
—24 and below	2.5
—23 to — 9	19.0
— 8 to + 8	53.5
+ 9 to +23	22.0
+24 and above	3.0

In other words, Differences ranging from minus 8 through zero to plus 8 are found in about 50 per cent of the cases and are, therefore, not significant. This is the condition found in the mass of school children at the present time, and without assuming this to be an ideal condition, or without assuming that this group is actually working up to its mental capacity, we must, nevertheless, accept the facts and from this basis of fact work toward a saving of the waste of intelligence that is apparent in the cases showing a difference of minus 9 or below.

The 21.5 per cent having Differences of minus 9 or below are obviously marked cases of wasted intelligence. They have ability to do much better and for some reason or other this ability is not being utilized. It is the business of the school to find out the reasons and

to stimulate them further. There may be many reasons, but, one reason, as we shall see in our second paper, is the very obvious one of misplacement in grade. Many of them are in a lower grade than they ought to be. Those showing Differences of minus 24 and below make up about 2.5 per cent of our cases and are, of course, the extreme examples of mal-adjustment. Prompt study of these cases is needed. They show the extreme of wasted intelligence.

Those showing Differences of plus 9 and above are children who are accomplishing more than is usually expected of children of their mentality. Again there are many reasons for this. One common reason is the pressure brought to bear on the slower child in order to make him keep pace with the grade suited to his chronological age, and the enormous amount of extra attention and especially skillful teaching that is lavished on this child. For these particular cases that show Differences of plus 9 and above, we believe that this emphasis is wrong. It is false economy at the present time when so much more intelligence is being allowed to run to waste.

The significance of the Difference ratings of schools will be somewhat different than that for individuals, because the Difference of a school is derived from the median mental and educational indices of all the children in the school. This decreases the spread of difference values. In all, 56 different schools were tested and the middle 50 per cent show differences from plus 3 to minus 3. For schools or classes a difference of medians above or below 3 becomes significant. All such schools having difference values greater than minus 3 show that educational attainment is too far below mental capacity and improvement should be expected.

The Grade Rating.—Our standardization of the educational tests has not been restricted to an age standardization merely. We have also prepared a grade standardization. This follows the scheme of the previous standardization, so that we may convert any educational score into an educational index according to grade. In doing this we compare the score obtained by a specific child with the other scores obtained by children of like grade and read off the index from the table. The indices have the same significance as before, namely, 0-19 meaning very poor work for that grade; 20-39 poor work; 40-59, average work; 60-79, good work; and 80-100, very good work.

In this way a teacher can compare children or classes with refer-

ence to their grade attainment. As we shall see, this educational grade index becomes important in deciding which children should be skipped or accelerated.

Mental or Educational Age.—If we use the scheme of indices as explained above there is, of course, no need to convert our results into mental or educational ages. Those who wish to do this, however, can readily do so by using the 50 index line in the index table. This gives the median scores for each age. Any particular score may thus be converted into a mental age. For example, if a child makes a score of 290 and if this is the median for age 11, the mental age will be 11. We may interpolate between ages and arrive at fractions of any age. In exactly the same way we may compute an educational age from the 50 index or percentile line of the table of educational scores.

Having these mental and educational ages, we can then in the usual manner calculate an intelligent quotient and an educational quotient, respectively. We may then use this I. Q. and E. Q. for any purpose we wish. The I. Q. divided by the E. Q. will give us a ratio, the significance of which will be similar to the Difference we have described above, only, of course, the values will cluster around 1.00, ranging upwards and downwards, instead of clustering around 0 with plus and minus quantities. The conversion of scores into mental and educational ages followed by the calculation of I. Q. and E. Q. and the ratio between them, involves much more work than the conversion of scores into indices and the simple subtraction of these indices. No advantage is gained by having ages and quotients, and we recommend the use of indices and differences on our tests.

The Reliability of the Tests.—One measure of the reliability of a test is the amount of correspondence between the rating of the same children tested by the same test at different times. The degree of reliability will be influenced by the homogeneity of the mentality of the group and by the time interval between the two testings. The less homogeneity the higher will be the correlation; the longer the time interval the lower the correlation.

The mental tests were given to a group of children in 1918 shortly after the tests had been devised. The same group were given the tests again in 1920. This group consists of 46 children ranging in chronological age from 10 to 14 in 1920, in mental indices from 28

to 81. Both the 1918 and 1920 tests were rated in accordance with the final standardization. What we are interested in knowing is to what extent the 1918 ratings are valid after two years have elapsed. Has a lapse of two years justified the ratings given in the first test? The coefficient of correlation between the two tests will denote the degree of correspondence in the ratings of the two testings. The correlation between the mental indices gives a coefficient of $r = .72$. In other words, the index that a child receives on the mental test is not likely to change much even after the lapse of two years.

If we rank the 46 cases according to their mental indices we find that the correspondence between them for the two testings is expressed by a rank coefficient of correlation of .77. In other words, we may say that on the whole a child does not vary greatly in rank after a lapse of two years. Some individual cases do show a great change in rank, while others maintain the same ranks, but on the whole the change is not great.

Of the 46 cases, 26 showed a gain in their mental index from 1918 to 1920, and the average gain for these 26 was 6.5 points; 18 showed a loss in mental index, and the average loss for these 18 was 5.8 points; while 2 had exactly the same index in both years. The greatest change in mental index was 20 points. This was made by a child who had a mental index of 27 in 1918 and 47 in 1920.

The correlation between the crude scores obtained on the tests will not be as great as that between the mental indices, and the longer the interval between any two testings the lower will the correlation for any given group become. This is due to the fact that children of different mental ability are progressing at different rates and the increase in crude score will vary according to the mental ability of the child. The Pearson Product Moment correlation for the crude scores for the 1918 and 1920 testings is 50 and the rank correlation is 58.

We may sum up this discussion of the reliability of the tests by saying that they show a remarkable degree of correspondence in the ratings of the children after a long period of time. This must not be interpreted as meaning that the index of any one child will remain constant over a long period. There are too many extraneous and accidental factors at work in group testing to warrant any such statement. Under good conditions, however, the chances seem to be

greatly in favor of rather close correspondence even after as long an interval as two years.

Agreement with Binet Ratings.—For 300 children, ranging from the 3rd to the 6th grade, a correlation of .47 was obtained between the mental index on the mental test and the I. Q. on the Stanford Binet. All these cases were tested during the same semester. This correlation between a 30 minute group examination and the thorough-going individual examination of the Stanford Binet is very satisfactory. We must bear in mind also the radical difference between the two tests, not only as to method of giving, e. g. group versus individual, but also as to content, e. g. non-language absolutely versus predominantly language responses.

The literary and scholastic nature of the Binet test seems reflected in the higher correlation between the indices obtained by the same group of children on the Educational Survey Test and the I. Q. on the Binet. In this case the coefficient is .52.

The correlation between absolute achievement on both tests as expressed by crude score on the non-language group test and mental age on the Binet is .80, which is very high and seems to show that the group test is discriminating well between different degrees of mental ability.

The correlation of the same group of cases between mental and educational indices is .62, which shows that on the whole educational attainment tends to go with mental ability, but that there is plenty of room for improvement, and therefore plenty of work for the psychologist to do in helping to adjust the child to the work he requires.

The correlation of the same group of cases between the Differences and the I. Q.s is .21. This shows as we shall demonstrate concretely in a later article, that we find children of all degrees of mental ability who are wasting their time in school and who are not working up to their real mental capacity.

Summary.—1) We have tried to show that the next logical step in psychological and educational measurement is the combination of mental and educational tests.

2) We have prepared two group tests, an educational and a mental, to measure school work and native ability respectively.

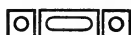
3) We have standardized these two tests and worked out a simple method for estimating the Difference between them.

4) This Difference is the most important value for school diagnosis, because it expresses the relationship between achievement and ability.

5) Untold waste in mental ability is brought to light by this method of study.

6) All intelligent school supervision will have to take account of the difference between ability and achievement and see to it that the educational achievement of each child is worthy of his ability.

DEPARTMENT FOR DISCUSSION OF RESEARCH PROBLEMS



Conducted by LAURA ZIRBES



This department has a two-fold function. It aims to serve research workers as well as educators, whose work brings them in close contact with children in the schools. It hopes to accomplish this service by suggesting research studies, which will meet well-defined school needs.

In order that this service may be real and effective, the co-operation of research workers and school people is desired. Correspondence with reference to the following questions will be considered in selecting topics for future discussions.

- a. Which of the studies proposed would help you to solve a practical problem?
- b. What topics might well be added to this list? Replies may be addressed to: Miss Laura Zirbes, 646 Park Ave., New York City.

A PROBLEM FOR CAREFUL RESEARCH:

HOW DOES THE ABILITY TO MAKE RAPID MENTAL ADJUSTMENTS AFFECT TEST SCORES AND PUPIL ACHIEVEMENTS?

The problem stated in the title is one which was suggested by a number of class room situations met during the past three years. A careful inquiry should shed light on a question which has many definite bearings on teaching practice and educational outcomes.

The problem first took definite form in connection with the use of Courtis practice tests in arithmetic in nineteen seventeen. It has since become significant in a number of connections with various types of material. A brief discussion of such instances will lead to a definition of the scope of the problem and to restatements in terms of concrete situations.

Instances from Arithmetic.—The Courtis practice materials were being used in grades IV, V and VI. A study of the errors made revealed a great number of those listed by Mr. Courtis in the handbook which accompanies his practice material. It revealed also mistakes which did not fall under any of the headings for which practice is provided in the test series. These mistakes were most frequent on Test A—Lesson Thirteen. This indicated that there was a factor in Test A which was not so active in other lessons. The errors of a great number of pupils are so similar that it seemed possible that they were all due to one cause. The errors in question

were of this sort. The subtraction examples which follow addition examples were worked as addition. The multiplication examples which follow the subtraction were worked either as addition or as subtraction. Tendencies to this sort of error were also noted in parts of examples, e. g. responses to 6×4 were 10 or 2. These errors were most frequent in the first few examples of each kind. There were some instances of pupils catching themselves in this type of error after a few examples had been worked. There were many cases in which the pupil could not see how the correct answer was possible until his mistake was pointed out.

A lesson was framed to give training in changing from one process to the other. This type of training is not provided in Courtis practice tests 1-12, inclusive, excepting as it occurs within the process of long division. The factor in question was played up in exaggerated form. Examples from Lessons 1-12, inclusive, were included. Time allowance was made to correspond with proportionate time allowed for examples in the respective lessons from which they were taken. Pupils took longer to work this lesson or did not finish in the prescribed time. The practice effects noted were increased accuracy and increased speed. Test A was then repeated with success at first trial. Courtis' suggestion is that pupils who fail to pass Test A need practice on Lessons 1-12. There was a strong indication that it is more economical to provide an extra practice lesson which would provide a type of training required by Lesson A but not supplied by Lessons 1-12.

The Woody arithmetic tests and Courtis Standard Research Test—Series B were given next. There were types of error in the Woody tests not found in the Courtis tests. Analysis showed this to be due to the grouping of similar forms in the Courtis test, and to the fact that in the Courtis Test the process could hardly be mistaken due to the character of the examples used and to the grouping. Each part of the Courtis Test, Series B could be worked with no reference to instructions as to the process required as each process is put in a separate test. This test then neglects absolutely a factor which is essential to success in Woody tests and in mixed process review lessons and tests. The Cleveland tests in arithmetic similarly ignore this factor, although the process cannot always be recognized by the form of the example. A mixed test was made composed of examples from the Cleveland spiral test. Both the

accuracy and speed scores were reduced by including different processes in one test.

These are the indications which suggest that it is necessary to train and test a skill factor which contributes to success, although it is not distinctly characterized as mathematical ability.

During the next year it became apparent that there was a similar situation and need for training due to the varieties of examples within each process. This made it necessary to supplement practice material provided by the Courtis tests in long division especially. For example, tests in which all the examples had zero difficulties were not difficult as soon as this fact was discovered. The same examples mixed with others in which there was no zero in the quotient showed errors which yielded to training with lessons which combined in one test, types of long division examples from various Courtis practice lessons. A type of example not provided in any of the tests was added because its entire omission gave the pupils a foreknowledge or check on accuracy not provided in other processes. The examples with remainders are the ones in question.

The detailed illustrations from the field of arithmetic make it possible to suggest more briefly the points at which a similar problem is presented in connection with other types of subject matter, and suggest investigation and experiment which may lead to solutions.

Illustrations from Reading.—A diagnostic study of the reading abilities of a fourth grade class was made during 1917-1918. Difficulties found were often due to a failure on the part of the child to set his mind for the particular reading task in hand, or to the persistence of one mental set through different types of reading. Concrete instances demonstrate the need of a type of training which will affect the ability to make mental adjustments rapidly and to see the necessity of change in mental set as it inheres in situations.

Pupils had very little silent reading experience. When the first measurement of oral and silent rate was made there was very little difference between the two scores for a great part of the class. Some of the differences between oral and silent reading were explained. Pupils were told that their lips might rest—silent reading is done by the eyes and the mind. Words are changed to pictures instead of sounds. Differences were illustrated, but no other in-

struction was given. Pupils were re-tested and there appeared variations which indicated that the mere consciousness of a need of change in mental set had an effect on rate of silent reading. Comprehension was checked so that the pupils were really reading at a faster rate silently, although the explanation made no mention of rate of reading. After this test, training designed to build up a *number* of silent reading "sets" rather than *one* was devised. Pupils were first given reading material which they were told to read with no further instructions as to the purpose of reading. They were then asked questions all of which could hardly be answered by a pupil who had maintained one "set" during the entire reading. The need of change was made inherent in the material, but the pupils did not see it.

The same material was repeated while pupils realized that their reading had not been effective. They were asked to check or underline parts of the material which they had not noticed before.

Next, a four part test involving change of set was devised. The first part was a very simple little story. The next was worded so that the sentences were not simple, although the words were not difficult. The next was an easy rhyme in which there was considerable repetition. The next was a problem. Each part was timed separately. There were some pupils whose rates in the four parts were nearly equal. The more effective readers judged by other comprehension and reproduction tests were those who varied their rate to suit the type of material. This indicated that there was a possible need of training in mental adjustment of this sort.

Analysis of the oral errors, which might have a bearing on comprehension, raised problems with regard to the possible defects of training which over-emphasizes one mental set. Some pupils disregarded word endings; others habitually read word by word without grouping into phrases; others skipped little words and went on regardless of meanings which were mutilated. Others were so intent on getting mental pictures and pleasurable experience that they paid no attention to precise words and often made interpretations which were not inherent in the text.

Brief Suggestions from Other Subjects.—It will be necessary to refer only briefly to the influence of mental set on spelling achievement. Pupils who spell well when spelling is uppermost in consciousness, become very inefficient spellers in circumstances which

demand the presence of other factors. Words learned for a given test or lesson can often not be recalled because the "set" in which they were acquired was not conducive to recall under other conditions.

Brief references to illustrations from geography illustrate the necessity for rapid mental adjustment in problem-solving wherever it occurs. For example, pupils were asked to outline the topics to be taken up in the order of their occurrence. Over half of them arranged topics with reference to another order. The preceding problem involved an arrangement of states as to their importance in a given industry. This "set" persisted. In another instance a pupil was asked to list the steps in a process. The work which had immediately preceded the question dealt with more general conceptions—namely, climate, location and soil condition. The answer given was in terms of the previous discussion.

Problems.—Some of the problems which grew out of these and similar experiences could profitably be made the subject of further study, and will, therefore, be stated in conclusion.

1. To what extent does "mental set" influence performance in standard subject tests?

2. Rapid mental adjustment is required by Binet Tests. If it is an intelligence factor which is affected by training as well as by age and development, how would general training in rapid mental adjustment affect Binet test scores?

3. To what extent must training of this ability be special training and to what extent does it transfer?

4. If drill devices are constructed from an analysis of functions into special habits, is it possible that an over training in special habits isolated for drill purposes contributes a mental "set" which must later be broken up?

5. What would lead pupils to put forth conscious effort in developing this ability?

6. To what extent can training involve a setting off of a group of reactions which is not peculiar to the special training itself, but which is a factor in other situations where that special habit must function?

PROBLEMS SUGGESTED FOR DISCUSSION IN
SUCCEEDING ISSUES.

How Do Attitudes Affect School Progress?

Indications of the insufficiency of intelligence tests as prognostic devices. The possibility of improving attitudes and the relative constancy of the intelligence quotient. Consequent importance of the study of the effect of attitudes on school results.

What Can Be Learned From the Manifestations of Nervous Exhaustion in School Children?

Possible causes. What school experiences are apt to draw heavily on nerve force? How can the nervous equilibrium be re-established and maintained? Individual differences. Factors beyond school control.

What Are the Situations in Which Reading Functions?

The scope of training in the intermediate grades. Present tendencies as noted in new texts and studies. The next step.

What Is the Significance of the Varied Types of Imagery Used in Problem Solving?

Problem suggested by one test given to groups differing in age and grade. Possible effects on curriculum and learning of individuals.

Lincoln School of Teachers College.

L. ZIRBES.

NOTES ON ARTICLES IN EDUCATIONAL PSYCHOLOGY IN CURRENT ISSUES OF OTHER MAGAZINES

Intelligence of Troops Infected With Hookworm vs. Those Not Infected. Garry C. Myers. Pedagogical Seminary, 1920, 27, 211-241. Those infected found to be of lower intelligence on the average than those not infected. The evidence does not prove that hookworm is the cause of inferior intelligence. Urges health education.

Intelligence at Senescence. M. F. Beeson. Journal Applied Psychology, 1920, December, 219-235. A study of the intelligence and education of inmates of public homes for the aged. An average I.Q. of 83 was found.

A Short Point Scale for Mental Measurement. Esther Reedy and James W. Bridges. Journal Applied Psychology, 1920, December, 258-262. Description of an abbreviated form of the Yerkes-Bridges Point Scale.

A Standardization and Weighting of Two Hundred Analogies. Rudolph Pintner and Samuel Renshaw. Journal Applied Psychology, 1920, December, 263-273.

Tentative Norms in the Rational Learning Test. Joseph Peterson. Journal Applied Psychology, 1920, December, 250-257.

The Accomplishment Quotient. Raymond Franzen. Teachers College Record, 1920, 21, 432-440. The theory and practice of basing school marks upon the quotient obtained by dividing the Educational Quotient by the Intelligence Quotient, thus displaying in terms of percentage the relation of actual progress to possible progress.

Table for Computing Mean Individual Scores in Educational Scales. Marvin J. Van Wagenen. Teachers College Record, 1920, 21, 441-451.

An Attempt to Measure the Comparative Importance of General Intelligence and Certain Character Traits in Contributing to Success in School. S. L. Pressey. Elementary School Journal, November, 1920, 220-229. Deals with health, "school attitude," preparation and ability as factors which condition success in school and success on a scale of intelligence.

Chicago Intelligence Test in Harrison Technical High School. O. Winter. School Review, December, 1920, 772-775. High School freshmen tested and grouped in teaching sections by the Chicago Group Intelligence Test. Moderate correlations shown between tests results and other measures of school ability like "marks."

Problem-Solving or Practice in Thinking (III). S. C. Parker. Elementary School Journal, November, 1920, 174-188. Third of a series of four articles

which consist of (1) account of the place of problem-solving in everyday life; (2) two sample lessons from the University of Chicago Elementary School; (3) "how great problem-solvers think."

Value of the Controlled Mental Summary As a Method of Studying. S. L. Pressey and L. W. Pressey. School and Society, November 27, 1920, 531-534.

Measuring the "Usefulness" of Tests in Solving School Problems. S. L. Pressey and L. W. Pressey. School and Society, November 27, 1920, 531-534.

The Educability Level of Five-Year-Old Children. G. G. Ide. The Psychological Clinic, May, 1920, 13, p. 146-173. An intensive investigation of the mental abilities of two groups of kindergarten children with prognostic suggestions. Clinical histories of many cases are given.

Some Volitional Patterns Revealed by the Will-Profile. June E. Downey. Journal Experimental Psychology, 1920, 3, 281-302. A method of measuring certain temperamental rates with which measures of intelligence may be supplemented.

Clinical Method As a Method in Experimental Education. Frank N. Freeman. Journal Applied Psychology, 1920, December, 126-142. Some reflections upon the relationship between group and individual study as a method of educational investigation; an illustration of the individual study of a case of "word blindness" and a statement of some of the advantages of such a method.

"The Attempt to Teach." Gladys E. Poole. Psychological Clinic, May, 1920, p. 173-190. Diagnosis by means of mental and educational tests and the method of clinical teaching of five cases of backwardness in reading.

Esthetics. Rudolph Pintner. Psychological Bulletin, 1920, 17, 331-335. A summary of articles appearing during the last year. Includes articles on aesthetic appreciation of children, measures of musical talent, etc.

The Psychology of the Thrill. Irving R. Kaiser. Pedagogical Seminary, 1920, 27, 243-281. A theory that emotions, instincts, interests, will, etc., are but means whereby thrills are sought and experienced; and which, in being experienced, afford us temporary relief from the feelings of stress which are brought about by the repression which civilization has placed upon our more primitive desires and actions.

Opportunities for College Graduates in Psychological Examining in Social Service Work and Education. Mildred Francis Baxter. Journal Applied Psychology, 1920, December, 207-219. A survey of positions now open to women trained in psychology, with an estimate of future possibilities.

Racial Differences in Mental Fatigue. Thos. M. Garth. Journal Applied Psychology, 1920, December, 235-245. Indians are less susceptible to mental fatigue as tested than whites, who are less susceptible than negroes.

Work, Fatigue and Inhibition. F. C. Dockeray. Psychological Bulletin, 1920, 17, 322-330. Summaries of articles on the psychology and physiology of work, etc., which have appeared during the past year.

NEW PUBLICATIONS IN EDUCATIONAL PSYCHOLOGY AND RELATED FIELDS OF EDUCATION*

I. BOOKS AND MONOGRAPHS IN APPLIED AND GENERAL PSYCHOLOGY.

The Psychology and Education of Subnormal Children: A Book for Teachers.—With the rapid extension of special classes for the education of subnormal children has developed an increasing demand for teachers who are specially trained both in the psychology and teaching of such children. A recent volume by Leta S. Hollingworth is written primarily for such teachers.¹ It gives an exceptionally good account of the “educational psychology” of subnormal children with the essentials of the physiological, medical and legal aspects of the subject. The causes of mental deficiency, the nature of individual differences, the means of identification, classification, definitions, the proportion of defectives among delinquents, prostitutes, unemployed, school laggards, etc., the physical traits, the nature of mental growth and maturity, special abilities and other features of feeble-mindedness are adequately discussed.

Two important chapters deal with the causes and prevention of mental deficiency; the effects of heredity, alcoholism, toxins, sensory defects, tonsils and adenoids, malnutrition, hookworm, malaria, injuries to nervous tissues, syphilis, glandular defects and the like. Speech defects, hysteria, dementia praecox and other nervous and mental disorders which may complicate mental deficiency are briefly considered.

The book emphasizes the educational aspects of subnormality. The defective is found to be more like the normal child in physical size and strength, in sensory acuity and in motor control, hence the emphasis upon sensori-motor instruction. The defective is found to resemble the normal child in instincts and emotions. His impulses are strong, but the capacity to acquire control of them, weak. With the feeble-minded, “moral training takes place only through the formation of specific habits.” However, the author makes it clear that the learning of the feeble-minded is not different in kind from the learning of the normal child of *the same mental age*. In general, the procedure is the same, the rate is the same,

*This department is conducted this month by A. I. Gates and H. O. Rugg.

¹Leta S. Hollingworth. *The Psychology of Subnormal Children*. New York: Macmillan, 1920. Pp. XIX + 288.

the amount of transfer the same. It is pointed out, however, that differences in emotional reaction, in habits, kind of information found in two children, one feeble-minded, one normal, both of mental age 8, are to be expected. Likewise their equal capacity to learn is only temporary since the growth, mentally, is different, hence the necessity for special classes or special schools, the history, organization and conduct of which are reviewed. The mental level ultimately reached by the defective is subnormal, so different content to some extent, and different educational and vocational aims than those of the customary school are demanded. The book includes what is useful for the teacher of special classes or for the teacher of such teachers. Each chapter is followed by an excellent list of references. The material throughout is well organized and skillfully written.

Books Dealing With Mental Disorders and Mental Hygiene.—Books dealing with mental disorders, their diagnosis and treatment, especially from various branches of the Freudian School, have been flowing from the press within recent years. Already psycho-analysis is being urged as a means of solving educational problems by applications of the method to children. Jelliffe (*The Technique of Psycho-analysis*, 1918) states that "a new science and application of pedagogy are being reared upon the data obtained by psycho-analysis." The theories underlying most of their systems are in most respects not in harmony with the accepted principles of experimental and systematic psychology, but psychologists have been slow to print thoroughgoing evaluations of these doctrines. Dunlap² has given us a timely and critical evaluation of the outstanding theories and practices of the Freud-Jung-Adler schools. Psycho-analysis is the "most important mystical movement of the nineteenth century" and "the most significant feature of mysticism which comes strangely to the surface in the Freudian school is its antipathy to experimentation."

Chapter I presents a brief history and description of mysticism with which psycho-analysis is identified. Chapter II gives an excellent exposition of the important postulates of the several systems, with evidence of the continued appearance of logical fallacies, the disregard of accumulated scientific data, and an almost

²Knight Dunlap. *Mysticism, Freudianism and Scientific Psychology*. St. Louis: C. V. Mosby Co., 1920. Pp. 173.

complete disregard of the recognized methods of scientific procedure and proof. In the third chapter the psycho-analysisists are given, it appears, some much needed first lessons in scientific methods followed by an exposition of some elementary facts concerning the nervous, muscular and glandular mechanisms, instincts, emotions, ideas, the laws of habit formation, etc. The book is apparently directed to the "general reader," but many pages will be specially useful to the student of education and psychology.

Destructive criticism is important, and just now much needed, in this field, but constructive contributions from the ranks of psychologists are even more to be desired. Hollingworth,³ as a result of his experiences as director of the psychological service of the U. S. A., General Hospital No. 30, at Plattsburg, has given us the most complete study of certain types of the functional neurosis which has appeared. The work is important because the examinations in most cases were elaborate and because probably two-thirds of the group of the 1200 soldier patients presented persistent psychoneurotic symptoms which were not severe enough to lead them to seek treatment under conditions of ordinary civil life. The book discloses the methods and tools employed by the psychological service in diagnosis and remedial treatment, together with a summary of statistical findings. For the first time, measures of the general intelligence level of such cases are provided, the average mental age being 11.7 years, nearly two and a half years below that of the average soldier. These age distributions show a striking parallel with educational attainments; 26 per cent received no schooling, but 15 per cent reached the 8th grade and but 4.4 per cent graduated from High School. When the cases were grouped according to symptoms: Group I, including those whose symptoms were *physical*, i. e., fits, tremors, stuttering, seizures, paralyses, sensory disturbances, blindness, cardiac troubles, etc.; Group III, those whose symptoms were definitely *psychic* or *subjective*, such as fears, worries, anxieties, obsessions, psychasthenia, nightmares, emotional disturbances, aboulias, depressions, etc.; and Group II, those suffering from both types of symptoms, and doubtful cases with such symptoms as headache, insomnia, chills, fatigue and fainting, a hierarchy of mental levels was found. For Group I,

³H. L. Hollingworth. *The Psychology of the Functional Neuroses*. New York: Appleton, 1920. Pp. XIII + 259.

the average mental age was 10.9 years; Group II, 12.0, and Group III, 14.5. Differences in specific symptoms are directly related to differences in mental level; the dullest displaying somatic disorders chiefly, those of slightly higher level displaying more strictly neurasthenic traits, while in the upper range are found the more exclusively mental and psychasthenic symptoms.

The book includes a wealth of statistical information from the use of the Woodworth Psychoneurotic Inventory, the change of symptoms following the armistice, the "scattering" of mental abilities, army rank distributions, occupational and geographical distributions, and evaluations of tests and methods. The data are analyzed and out of them has developed a theory of the mechanisms involved in psychoneurotic conditions. It is the simple and well known mechanism of associative shifting or the "redintegrative mechanism" which has long been fundamental in our explanations of perception and other integrated reactions. ABCD, etc., elements of a complex situation, some or all of them, produce a complex response XYZ. Later the common element C, met among other stimuli, F.G.H., "redintegrates" XYZ and later H, associated with C in the second case, appearing alone with J.K.L. redintegrates the response again. The theory is discussed in detail and that it may be readily substituted for such mystical or pictorial concepts as "sphoning of the affect," "repression into the unconscious," "transfer of libido," etc., is suggested. The explanation here offered is simple enough to satisfy the rigid requirements of the law of parsimony; it merits careful consideration and test in the hands of professional workers. It promises to be the most important theoretical contribution in this field within recent years.

Significant details of two years' experience as an independent clinical psychologist are given in a brief account by Martin,⁴ formerly Professor of Psychology at Stanford University. An outline of methods of diagnosis and treatment, largely educational, with representative case studies is given. The book contains little new theory, practice being based on the principles of the Freudian School and indirectly on the methods of experimental psychology, with a rather unusual reliance upon transfer effects. The psychologist may doubt the value of many of the remedial exercises that are suggested despite the appearance of cures. As Janet reminds

⁴Lillian J. Martin. *Mental Hygiene*. Baltimore: Warwick and York, 1920. Pp. VI + 89.

us, "The temple of Æsculapius has cured thousands of patients; Lourdes has cured thousands of patients; animal magnetism has cured thousands of patients; Christian Science has cured thousands of patients—and psycho-analysis has cured thousands of patients.—It is extremely difficult to eliminate other influences which may have modified the disease." Suggestion, stimulation, encouragement, sympathy, ridicule or other incidents within or outside the control of the practitioner, rather than the pictures, the gymnastics, the music or practice on control of imagery may have been the cause.

A conventional classification and statement of principles of causation and treatment of mental disorders will be found in a recent syllabus prepared by Charles B. Thompson.⁵

A New Manual of Mental Tests.—The scientific movement in education would be much enhanced if "directors of research," and designers of intelligence and educational tests would apply the type of careful procedure and cautious interpretation which the authors of the new Dewey-Child-Ruml test manual have employed. There is a good deal of reason to believe that we "scientific educationalists" are developing a rather crude and uncritical procedure—especially so, in our hasty publication of tests and test results; in our slipshod making and testing of hypotheses (many of us never bother to make them in our investigations and studies) and in our naive interpretation of "correlations," "probable errors," forms of distribution and the like. This new book⁶ will serve as a fine stimulus to more careful work in the testing field. These workers, leaders in a wider co-operative group, have tried out, assembled into a working unit, and have standardized in a preliminary way a number of new mental tests, together with groups of tests already in current use. Their purpose was to discover if an individual analysis (of normal pupils) can be made *by* tests, not *from* them: an analysis which is definitely the outcome of purely objective test results. In addition to using such available tests as the Binet, the Yerkes-Bridges Point Scale, the Woolley Identification of Forms, the Healey Picture Completion test and the like, the group (especially Miss Dewey and Professor Hayes) have devised a number of new tests. These include such examples as: cart construction, narrative pictures, needle threading, nail driving, etc.

⁵Charles B. Thompson. *Mental Disorders*. Baltimore: Warwick and York, 1920. Pp. 48.

⁶Dewey, E., Child, E., and Ruml, B. *Methods and Results of Testing School Children*. New York: E. P. Dutton & Company, 1920. Pp. XII + 176.

The testing was done with individual children, the time of testing varying from five to six periods, from twenty-five to forty-five minutes each. From these tests an abbreviated maturity scale was prepared, one for boys, another for girls. The one for boys consisted of seven tests; that for girls of six tests. In making this scale the writers avoided the assumptions which underly the different age-grade scales in current use—namely, “that a test situation gets at the essential factors in general intelligence and that tests highly correlated with their own total were necessary functions in identifying intelligence.” They have so selected their tests that they do not overlap, and with reference to a measurable criterion *outside the test series*. They have stated the scores in terms of relation of each child’s deviation from the average to the standard deviation of his age group.

Test scores and norms are reported in detail, together with diagrams showing the complete distribution of scores made on each test. The statistical methods employed are unique for such publications. For the first time definite use has been made of the regression equation in stating scores and norms. They report the regression of “scores on age” instead of using the commoner method of giving the arithmetic mean (or median). They did this so that erratic effects obtained by sampling would be neutralized.

We have held up this manual as a useful model of scientific precision in formulating and standardizing tests. (Its cautious interpretations also set an example in the treatment of test material.) We would emphasize the fact that this is a useful example of scientific technique and *nothing more*. We would not have it thought that we believe that the procedure and results can be employed in public schools today. They cannot. They are administratively impossible. At the present time probably not twenty-five educationalists in America can interpret and use its procedure and results correctly without further careful statistical training. It is possible, however, to obtain from this little book (impracticable as it may be today) an object lesson for the improvement of our investigational technique and of our own mental attitudes toward rigorous scientific work in education.

The Eyesight of School Children. A treatment of the problem of vision in the schools of an authoritative character and more complete than the conventional chapter in books on hygiene, is needed.

Berkowitz⁷ has fulfilled this need by presenting a summary of facts concerning prevalent eye defects with a series of recommendations concerning diagnosis and treatment which have been approved by a committee of leading oculists and ophthalmologists. The relation of day and artificial lighting systems, the arrangement of desks, interior decorations, blackboards, etc., to the hygiene of vision are samples of problems which are treated.

II. BOOKS AND REPORTS IN EDUCATION.

Two New Books on the Elementary School Curriculum.—Two books have appeared recently which deal with the reorganization of the elementary school curriculum.^{8, 9} The writer of each one has spent some years in the reorganization of school curricula. Each writer is applying the best there is in the James-Dewey-Thorndike pragmatic and social philosophy and psychology. While both books plead for "reorganization," yet the two are nearly as unlike as it is possible for them to be.

Professor Bonser agrees with Professor Meriam that it is "ideally desirable" to *organize a curriculum about activities* rather than subjects; that people's activities shall be so utilized that "the curriculum shall directly emerge from them and hence, fit for them." The curriculum emerges as "a series of purposeful activities," and is the "means of providing them." Each of these two books aims to do what our pragmatic and social philosophy has been preaching for a generation: that is, construct a curriculum completely on the criterion of social worth. "Only as the knowledge, habits and attitudes, and appreciations developed in the school are operative in meeting the problems of life are they of any worth." To this statement they will both subscribe.

Striking differences appear in the applications which the two men make of their theories. While it is "ideally desirable," yet Professor Bonser thinks it is "practically" impossible to completely organize a curriculum about "activities." His method is to lead school men gradually to change their procedure. So we find, in his book, a discussion of all the established school subjects. His device

⁷J. H. Berkowitz. *The Eyesight of School Children*. Washington; Govt. Printing Office, Bureau of Education, Bulletin No. 65, 1920. Pp. V + 128.

⁸Bonser, F. G. *The Elementary School Curriculum*. New York: Macmillan, 1920. Pp. XVI + 446.

⁹Meriam, J. L., *Child Life and the Curriculum*. Yonkers: World Book Co., 1920. Pp. XII + 533.

for effecting change gradually is the "project." "Changes must be gradual, not abrupt for them to be adopted by teachers," untrained and yet habituated as they are to a "subject" basis of organization. He recommends a "natural transition," however, through the pages of his book. "Beginnings may be made with single units," this means blocks of material organized as problem-projects.

Professor Meriam's book is, first, a more vigorous denunciation of the existing curriculum. Several chapters are devoted to this sort of argument and to developing principles of curriculum-making.

When we come to his "contents of a curriculum" we find that Professor Meriam has indeed had the courage of his convictions and has completely broken away from the established school subjects.

He sets up four school studies in his school (The University Elementary School, Columbia, Mo.): 1—Observation; 2—Play; 3—Stories; 4—Handwork. Here the program is *not* organized as reading, writing and arithmetic; history, geography, science, civics are all excluded from his curriculum scheme. To be sure the material of these "studies" is all classifiable under one or more of the established subjects. Quite clearly the writer confuses *general* activities (like play) with devices for stimulating such (like stories), and in turn with mental functions (like observation) by discussing these as though they were coordinate. Observation is to form the core of the work in social and industrial activities of the intermediate grades, animal life, earth and sky and people in the primary grades; local industries in grades III and IV. Stories including music, poetry and pictures is the one phase of the four-fold scheme that is to train primarily for the *leisure* life, according to the writer. The burden of his argument is, first, that by organizing "reading" as an incidental school activity reading becomes enjoyable and that the amount read is large. No definite measurements are reported to show to what extent the children in the school actually *can* read, however. This is *true* throughout the book.

The book is full of inconsistencies and contradictions; for example, its frequent questioning of the worth of quantitative methods, while subsequently calling to the author's support just such methods. It is unordered, very repetitious, and it appears to the reviewer, needlessly argumentative and exhortatory. It is clear that it could not be used as a text for *prospective* teachers—it would confuse rather than help. For the experienced and capable teacher,

for trained faculties of normal schools and city school systems and for graduate students in education, we believe Professor Meriam's book will serve as a "stimulating irritant" to the discussion of the mooted questions of curriculum-making.

Professor Bonser's book on the other hand will serve as an admirable text book for prospective teachers or for reading-circle work. It is orderly, systematic, carefully written; will introduce students and teachers to modern problems and methods of organizing a curriculum and will show how the traditionalism of current practice may gradually be broken through. It will train teachers well for present jobs. The present reviewer doubts if it will serve to stimulate as much thoroughgoing discussion of the *reconstruction* of the curriculum, however, as will "The Child and the Curriculum," even though the latter may seem to some of us a definite example of what *not* to do.

A "Text"-Book on the Supervision of Instruction.—Books on the supervision of teaching have rarely contributed to improve the actual class-room work of teachers. The chief reason for this is that they have been compilations of platitudes about the purposes of teaching, sets of principles of "methods," general directions, etc. While Professor Nutt's new book¹⁰ is an improvement on the mass of such books it cannot be said to set a really new mark in this field. It continues the precedent set by the earlier books, of giving many chapters of "Principles Underlying the Supervision of Instruction" (Chs. III, IV, V and VI), and scores of pages of "devices of supervision," which are largely *comment about devices* and not the devices themselves. Occasionally one meets concrete *examples* of supervision (e. g. ps. 158-59); now and then a "device" is presented in sufficient detail to inspire a supervisor to go and do likewise. One reads the pages of this book, however, with dismay at the accumulation of principles and methods and with an unsatisfied sense of not having seen and learned how to use the devices and methods which good supervision must be employing. Its chief function will be to serve as a guide for college class discussions in which the instructor supplies the rich background that is needed for understanding the "texts" of the book.

Its principles of method and its survey of devices are in sound accord with our current psychology. The suggestions for the actual

¹⁰Nutt, H. W. *The Supervision of Instruction*. Boston: Houghton Mifflin Company, 1920. Pp. XI + 277. \$1.80.

supervision of class work are sensible and helpful. Its great deficiency is in omitting a wealth of pertinent *examples* for each phase of its discussion.

The Need of Federal Assistance for the Nation's Schools.—Since “the entire schooling of the average native-born citizen has cost the public less than one hundred and fifty dollars—an amount of comparable perhaps with that which the village grocer spends on his daughter’s piano lessons,” and since it is found that in five prosperous Middle-Western States the machinist, plumber, bricklayer and blacksmith are paid more than twice the average salary of elementary teachers, Keith and Bagley,¹¹ for these and other reasons, have advocated the creation of a Department of Education provided with means by the Federal Government for remedying the situation. Aside from a comprehensive statistical study of the existing conditions among schools with an effort to trace them to their causes, the book outlines the historical development of the policy of Federal aid, eventuating in a critical consideration of the Smith-Towner Bill, as representing the most satisfactory current proposal. Although the book is written in form suitable for the general reader, it cannot be too strongly emphasized that the professional man in education and psychology should seek intimate acquaintance with material such as is here assembled.

A Committee Report on the Reorganization of Science in Secondary Schools.—After more than seven years of discussion and experimentation, the Committee on Reorganization of Science in the Secondary Schools, under the leadership of Professor Caldwell,¹² has given a report which is a marked improvement upon the conventional “writ of opinion” variety. Although no experimental data are given, the suggestions with regard to materials and methods are offered after having been successfully tried in many schools. Projects, of a sane and useful sort, carried out under careful guidance by the teacher is the foundation of the method. The relation of projects to topics, “the natural way of working,” choice of topics and materials, laboratory and classroom procedure, demonstrations, excursions and clubs are features which are considered in detail. While this sort of work may be in no sense considered the equivalent

¹¹John A. H. Keith and William C. Bagley. *The Nation and the Schools*. New York: Macmillan, 1920. Pp. XVII — 364.

¹²Otis W. Caldwell and others. *Reorganization of Science in Secondary Schools*. Washington: Govt. Printing Office, Bureau of Education, Bulletin No. 26, 1920. Pp. 62.

of experimental research and experimental teaching, it marks a distinct improvement in committee procedure.

III. ADDITIONAL PUBLICATIONS RECEIVED.

A. PUBLICATIONS IN GENERAL AND APPLIED PSYCHOLOGY.

- EDMAN, I. *Human Traits and Their Social Significance*. Boston: Houghton Mifflin Company, 1920. Pp. XI + 467.
- GODIN, P. *Growth During School Age*. Boston: Badger, 1920. Pp. 268. Translated by S. L. Eby.
- LAY, W. *Man's Unconscious Passion*. New York: Dodd, Mead & Co., 1920. Pp. V + 245.
- MUSCIO, B. *Lectures on Industrial Psychology*. London and New York: George Routledge & Sons, Ltd., and E. P. Dutton & Co., 1920. Pp. 1 + 300.
- SCHOFIELD, A. T. *The Mind of a Woman*. New York: E. P. Dutton & Co., 1919. Pp. VII + 120.
- SEASHORE, C. E. *A Survey of Musical Talent in the Public Schools*. Iowa City: The University, Iowa City, Ia., 1920. Pp. 7 + 36.
- TERMAN, L. M. *Condensed Guide for the Stanford Revision of the Binet-Simon Intelligence Tests*. Boston: Houghton Mifflin Company, 1920. Pp. 7 + 32.
- WILLIAMS, J. H. *A Survey of Pupils in the Schools of Bakersfield, California*. Whittier, Cal.: Whittier State School, Department of Printing Instruction, 1920. Pp. 1 + 43.

B. PUBLICATIONS IN THE GENERAL EDUCATIONAL FIELD.

- BRIGGS, T. H. *The Junior High School*. Boston: Houghton Mifflin Company, 1920. Pp. IX + 350.
- CUBBERLEY, E. P. *The History of Education*. Boston: Houghton Mifflin Company, 1920. Pp. XI + 849.
- CUBBERLEY, E. P. *Readings in the History of Education*. Boston: Houghton Mifflin Company, 1920. Pp. IX + 684.
- HANUS, P. H. *School Administration and School Reports*. Boston: Houghton Mifflin Company, 1920. Pp. III + 200.
- SHOWALTER, N. D. *A Handbook for Rural School Officers*. Boston: Houghton Mifflin Company, 1920. Pp. IX + 213.
- STOCKTON, J. L. *Project Work in Education*. Boston: Houghton Mifflin Company, 1920. Pp. V + 166.

C. MISCELLANEOUS PUBLICATIONS.

- FINDLAY, J. J. *An Introduction to Sociology*. London: Longmans, Green & Co., 1920. Pp. IX + 304.
- LAMBERTSON, F. W. *The Rules of the Game*. New York: The Abingdon Press, 1920. Pp. 5 + 208.
- LAMBERTSON, F. W. *The Rules of the Game. Teachers' Manual*. New York: The Abingdon Press, 1920. Pp. 5 + 77.
- PARKER, D. H. *The Principles of Aesthetics*. Boston: Silver, Burdett & Co., 1920. Pp. V + 374.

THE JOURNAL OF EDUCATIONAL PSYCHOLOGY

Vol. XII

FEBRUARY, 1921

No. 2

EDUCATIONAL PSYCHOLOGY AT THE CHICAGO MEETINGS OF SCIENTIFIC SOCIETIES

ARTHUR I. GATES,
Teachers College, Columbia University.

The American Psychological Association, Section I (Psychology) and Section Q (Education), of the American Association for the Advancement of Science held their annual meetings jointly at the University of Chicago, December 28-30, inclusive. The attendance was unusually large and the programs were more than normally extensive. Some notion of the interest displayed in educational psychology is given by a classification of the papers offered. Aside from the addresses of retiring officers, eighty papers were read, and of these one-half were either studies of educational problems or studies which could be directly applied to education. Twenty-seven titles dealt in particular with tests; general psychology claimed eight titles, experimental nineteen, comparative four, social four, clinical nine, and industrial six.

V. A. C. Henman, University of Wisconsin, retiring vice-president of Section Q, in an address, "The Measurement of Intelligence," reviewed past accomplishments and evaluated the present status of testing. "The conspicuous development in the application of scientific method to educational problems during the past two years has been the construction and use of intelligence tests in schools and colleges." "What we have learned about the influence of environment on mental traits and the failure of environment to alter them materially; what we are now learning about the constancy of the intelligence quotients, and the fact that mental alertness is given like retentiveness once for all with one's native constitution, magni-

fies the function of the school in selecting individuals and minimizes its function in training them." The teacher is to become a "diagnostician and director" rather than a "trainer." Classification by mental age and "sectioning" on the basis of the I. Q. were urged. However, "tests are greatly needed of ability to deal with things and other human beings, what Thorndike calls mechanical and social intelligence." Finally, the misuse of tests in schools and elsewhere and the present tendency toward multiplication of group tests, partly due to the urge of publishers, were criticised. What we need is not more tests, but refinement, standardization, and analysis of a few.

C. H. Judd, speaking at a symposium in "Problems of Psychology," criticises the current practices of mental and educational testers who are often "content to let tests be tests, and linger on the outskirts of real science." "The practices of testing are running far ahead of the explanatory science." The necessity of refinement of tests was urged, and likewise the need of analytical work. A test in reading, e. g. is made, a median and coefficient of correlation given, but the tester, and above all, the teacher has but a vague idea of what it all means. It was urged that more adequate preparation in experimental and scientific technique be acquired by students, that the limitations of statistical procedure be made clear, that more interest be centered on analytical problems, and, that most of all, the development of the explanatory science should be furthered.

Views, more or less in harmony with those of Judd, were expressed in open discussion by Raymond Dodge, Robert M. Yerkes, Joseph Jastrow, W. S. Hunter, and others, and it became apparent early in the meetings that the relation of the practice to the science of psychology, particularly in connection with tests, would be the outstanding discussion of the convention.

F. N. Freeman, University of Chicago, in an important paper, took up our lack of information concerning the curve of mental growth, and the questionable practice of estimating and applying the 'Intelligence Quotient' from point scales. For the I. Q. to be valid, the rate of growth must follow a logarithmic curve, producing a constantly increasing amount of overlapping from year to year or a straight line curve, which likewise produces increased overlapping. An examination of the results of recent group intelligence tests indicated that the use of the intelligence quotient was a very question-

able procedure. The need of empirical work in the technique of test construction and interpretation was emphasized.

M. E. Haggerty, University of Minnesota, in a paper, "The Essential Criteria of Mental and Educational Tests," urged a more careful evaluation of the units of "battery" tests, and the publication of detailed statistics upon which tests are based. Tests are thrown upon the market without proper standardization; at least, the data for the test units are not made available. Critical study of certain tests shows parts or the whole of certain tests to be of doubtful value. The publication of both age and grade norms was advocated.

Rudolf Pintner, Ohio State University, illustrated a frequently misleading practice of measuring educational attainments in schools, basing recommendations upon them when the intellectual status of the pupils is not known. Classes and schools differ very greatly in capacity to profit by instruction; measures of attainment are meaningless unless supplemented by intelligence scores. A group test, measuring both general intelligence and school attainments, which may be given in a short time with results sufficiently accurate for school surveys was described.

Throughout the meetings there was an active interest in defining more exactly the uses, and especially the limitations, of numerous educational and mental tests. The need of more careful evaluations of the existing tests, more satisfactory standardization and more cautious interpretation of results was constantly emphasized. That intelligence ratings are more useful when supplemented by measures of other traits, physiological, emotional, temperamental, and that these features may yield to quantitative treatment by practicable methods was indicated by the contents of several papers.

Bird T. Baldwin, University of Iowa, found it useful to supplement intelligence ratings with measures of certain anatomical traits which give an index of physiological maturity. Physiological age has a direct bearing upon emotional traits, social relations, and pedagogical advancement. Questions of rapid promotion, it was urged, should be settled on the basis of physiological maturity as well as mental maturity. In general, there is a good correlation between physiological age and mental age, but occasionally the disparity is significantly great. Height, weight, and radiographs of the carpal and metacarpal bones are indicative of physiological maturity.

June E. Downey, University of Wyoming, devised a group test, including many features of the original individual test, of the will-temperament, for use with students from the seventh grade through college. The group test is less valuable than the earlier for refined results, but yields certain information concerning motor impulsiveness, co-ordination of impulses, etc., which is a useful supplement to an intelligence rating. Patients suffering from various mental disorders present characteristic will-profiles. M. J. Ream, Carnegie Institute of Technology, constructed a group test using many of the Downey devices which gave fair correlations with success in salesmanship.

A scheme for describing by contrasts, the interests of an individual, was set forth by J. B. Miner, Carnegie Institute of Technology. Standard psychographs for both sexes, and for students in different types of courses, were shown. The plan was proposed as useful in helping the youth to discover the vocation which approximate his complex of interests.

Results from the Use of Tests.— J. W. Bridges, Ohio State University, found correlations averaging $+0.36$ between academic grades and point scores on the Army Alpha with a group of nearly 6000 college students. Variations of correlations with sub-tests indicated that there should be specific tests for students of different colleges; engineering, etc. William F. Book, Indiana University, in a study of over 6000 high school seniors, found that approximately the same percentage of dull, average, and bright students intended to go to college. Twenty-two per cent. of the brightest students expected not to go to college. Dull students preferred professional or technical college courses, the bright desired liberal arts more frequently. James P. Porter, Clark University, suggested certain methods of analyzing the mental aptitudes and inaptitudes of adults by detailed examination of relative success in sub-tests of the Otis and Thorndike series. L. L. Thurstone, Carnegie Institute of Technology, found the predictive value of his thirty-minute intelligence test for engineering students to be equal to the records of scholarship for four years in high school. The predictive value of the test was higher in liberal arts colleges and normal schools than in engineering schools. J. E. W. Wallin, St. Louis, finds the judgments of teachers about as reliable as a group intelligence test (Pressey Primer) for selecting mental defectives; the Stanford-

Binet being more reliable by far than either. H. H. Goddard discussed various features of abnormal functioning of the mind, upon which little light is thrown by intelligence tests.

Lewis M. Terman, Stanford University, reported on the progress of a study of nearly 100 children of superior ability. Ratings on forty-six mental, moral, and physical traits, as well as intelligence measures, are being secured annually. Results so far indicate a constancy of the I. Q. with a few changes of considerable magnitude. Ratings on social adaptability show a tendency to improve as the child matures. Some details concerning Betty F., a gifted juvenile author, I. Q. 188, who had not attended school at eight, but had read 750 books and written enough verse and prose to fill a fairly large volume, were given.

President Walter Dill Scott, Northwestern University, speaking on "Psychology in Industry," urged the development of vocational tests for the placement or selection of men rather than the mere elimination of the unfit. Data were presented showing the influence of intelligence level upon contentment with different jobs. With a certain optimum and intelligence, the man "likes" the job, whereas a man of higher intelligence finds it monotonous, and a lower intelligence finds it too exacting.

Walter S. Hunter, Kansas University, measured 715 American Indians of full and partial breed, finding a median Otis score of 82.6, as compared to a score of 122.6 for 1366 native whites of equal age. With comparable groups, it was found that intelligence varied positively with degrees of white blood possessed by the Indians. Thomas R. Garth, University of Texas, in tests of 371 Indians, found a similar relation of intelligence to blood, but was unable to entirely eliminate variations in educational advantages among the groups.

S. A. Courtis, Bureau of Statistics, Detroit, presented the results of an age-grade survey of Detroit schools, in which children were classified by nationalities, grades, and schools. Retardation rises to a peak at the fifth grade, falling to a minimum at the eighth. Acceleration increases gradually to the eleventh grade. On the whole, the greatest retardation is found among negroes and Italians; the greatest acceleration among United States whites, Scotch, and Russians.

Ada Arlitt, Bryn Mawr College, presented data showing that native whites surpass slightly in Stanford I. Q.'s Italians and

negroes of similar social status. Dividing all subjects into five groups according to social status, the I. Q.'s of each average 125.9, 118.7, 106.5, 87.0, and 83.4. Karl T. Waugh, Berea College, presented a comparison of Oriental and American student intelligence ratings which could not be attributed securely to racial differences.

Buford Johnson, Johns Hopkins University, found that children from 8 to 20 per cent. under weight, according to present norms, perform as well in mental functions as children of normal weight when other conditions, especially intelligence level, are equal. Children 8 or 9 per cent. below the norm in weight grow at standard rate in weight, whereas those from 10 to 20 per cent. below the norm show less than the expected gain.

New Educational and Mental Tests.—A group intelligence scale for the primary grades, pictorial in character and involving no reading, writing, or complicated drawing, was devised and standardized tentatively by Forrest A. Kingsbury, University of Chicago. The test yielded an average correlation coefficient of $+0.69$, with the Stanford-Binet Mental age of 1300 primary pupils. A convenient, combined mental-educational test, for survey work, has been devised by Rudolf Pintner, Ohio State University, and standardized on 4300 cases. It affords a measure of the mental status of a class, and provides a convenient means of evaluating educational attainments in terms of innate ability to profit by instruction. Sidney L. Pressey, Indiana University, presented a new scale for measuring attainments at the end of the 7th grade, in history, reading vocabulary, English composition and arithmetical reasoning. English composition was measured by the novel method of scoring the accuracy of children's judgments of compositions of known value.

Experimental and Statistical Studies of School Subjects.—Shepherd Ivory Franz, in the presidential address, discussed the need of work in analytical psychology in the interests of psycho-pathology, and for the solution of problems concerning the relation of mind and body. The brain pathologists and the histologists are hampered by lack of information concerning the structure and correlation of mental states, e. g., the nature and function of imagery, of perception, etc. A review of recent cases of abnormalities in behavior with the corresponding brain lesions was offered, illustrating, in particular, the need of a more useful analysis of mental activities. In the first session on educational psychology, a similar need

of analytical work to define the mental processes involved in school functions was emphasized, both in formal papers and in informal discussion. That the mental processes involved in school subjects may be profitably attacked with the technique of the laboratory was illustrated in several papers. A movement toward a return to analytical studies by experimental methods would be heartily approved by many.

G. T. Buswell, University of Chicago, reported a commendable study of oral reading. By an ingenious method graphic records of the positions of eye and voice in reading were obtained. The better reader has a better eye-voice span, which is characterized by several variations from that of the poor reader. Such analytical studies pave the way for an effective method of instruction in reading. W. S. Gray, University of Chicago, gave details of an analytical diagnosis of a case of backwardness in reading, in which the difficulty was discovered and remedied by appropriate training. Paul West, University of Chicago, by the use of photographic apparatus, determined certain characteristics and conditions of motor rhythm in writing movements. Essentially speed and rhythm are determined by the radial length of the arc inscribed in writing. Certain rhythmic drills in the case of children backward in writing were found to be useful.

Drawing ability of 725 children, determined by the use of standardized scales, was found to yield a correlation of 0.83 with tests of visual memory, and a correlation of 0.82 with tests of perception of perspective, devised by Elmer E. Jones, Northwestern University. The two tests gave a coefficient of 0.69.

The results of measurements of spelling, in the Virginia Survey, were presented by Frederick S. Breed, University of Chicago. Rural schools were from 2.7-3.8 years, and city schools, 2.0-3.2 years below the achievement of representative American eighth grade systems.

Luella W. Pressey, Indiana University, devised four tests of silent reading, two based on general press literature, one on scientific passages, and one on poetry. It was found that the tests on general matter gave a correlation of 0.85, but that the coefficients with reading poetry were 0.38 and 0.31, respectively, and 0.35 and 0.49 with scientific material. Proficiency in reading poetry and scientific material is correlated $+0.56$. "It suggests that 'silent

reading ability' varies with the type of matter used; and, that if may (after freedom from oral reading habits has been achieved) be a composite of general ability, background information in the subject dealt with, and interest. Measurement of 'silent reading ability' in the upper grades, or teaching directly toward the improvement of 'silent reading ability' as an entity, would hardly seem warranted under such circumstances."

E. K. Strong, Carnegie Institute of Technology, in his address as retiring vice-president of Section I, urged that courses in vocational education be built, not upon opinions of university men or industrial executives as heretofore, but upon the results of actual job analyses. An outline was given of the technique of job analysis which consists essentially of discovering: (1) what an executive does; (2) what he must know in order to do it; (3) other abilities valuable but not necessary; (4) the progress of successful workers.

Studies of Learning in Laboratory Functions.—Joseph Peterson, George Peabody College, in studies of maze learning by human subjects, found that "learning seems to be possible only by the overlapping of the effects of successive stimuli, in accordance with the principle of "completeness of response," when frequency and recency factors are made negative. John F. Shepard, University of Michigan, described new forms of maze apparatus for human and animal subjects. Katherine Ludgate, University of Chicago, found that guidance of the learner's hand through a maze in the first attempts shortened the total time required to learn.

Distributed practice in memorizing digits is more fruitful, as a rule, than concentrated learning, according to experiments of Edward S. Robinson of Chicago University. Certain sex differences, generally in favor of men, in the case of mirror tracing, appeared in a study made by George S. Snoddy, University of Utah. A detailed study of interference effects of card sorting, pencil mazes, etc., was reported by J. F. Dashiell, University of North Carolina. Stevenson Smith, University of Washington, discussed the mechanism of trial and error learning of animals. L. A. Pechstein, University of Rochester, found, with both human and animal learners, that massed learning, when the maze is new and short, is to be preferred to distributed effort. In certain cases, long problems were best learned by massed attack upon parts.

Augusta F. Bronner, Judge Baker Foundation, has devised a series of learning tests for practice work which involve functions comparable to school subjects. That school children are woefully ignorant of efficient methods of studying was the finding of an investigation by A. S. Edwards, University of Georgia.

Business Transactions.—Of importance to educators and psychologists alike was the report of Bird T. Baldwin, chairman of the Committee on Qualifications and Certification of Consulting Psychologists. The report, approved by the Clinical Section of the A. P. A., and by the Association as a whole, provides for the standardization of academic and professional equipment of clinical psychologists, for examinations, and for a written testimonial to be given to successful candidates. The action was taken as a means of protecting the public from ill-equipped practitioners until the States shall provide boards to pass upon the fitness of professional workers in psychology and shall issue certificates. A model bill to assist legislation in the States was drawn up by the committee.

C. E. Seashore, University of Iowa, reported for the committee on a journal of abstracts of psychological literature. Final action was not taken, but it is likely that the "Psychological Bulletin" will be converted into a six-issue journal, containing abstracts of the literature of all languages. At present, the technical literature of psychology is too widely scattered to collect, too expensive to purchase, and too voluminous to read in the original. An organ to serve as a clearing house of literature, especially that of foreign origin, is very much needed.

The following are officers-elect for the American Psychological Association for 1921:

President, Margaret Floy Washburn, Vassar College. Members of Council—G. F. Arps, Ohio University, and W. S. Hunter, University of Kansas. G. M. Stratten, University of California, and W. B. Pillsbury, University of Michigan, were elected representatives to the Natural Research Council. E. K. Strong, Carnegie Institute of Technology, was elected representative to the Council of the A. A. A. S. G. M. Whipple, University of Michigan, was elected chairman of Section I (Education) of the A. A. A. S. Thirty-nine candidates were admitted to membership in the A. P. A. The A. A. A. S. will meet in Toronto in 1921, but the place of meeting of the A. P. A. has not been selected.

A SURVEY OF THE THREE FIRST GRADES OF THE HORACE MANN SCHOOL BY MEANS OF PSYCHOLOGICAL TESTS AND TEACHERS' ESTIMATES, AND A STATISTICAL EVALUATION OF THE MEASURES EMPLOYED.

CLARA F. CHASELL,

Psychologist of the Horace Mann School, Teachers College, Columbia University.

and

LAURA M. CHASELL,

Instructor in Psychology, Ohio State University.

The practice of making chronological age almost the only basis of admission to the first grade, and the tendency to promote by age rather than ability, are the two factors which help most to forge the chains of our fatal lockstep system of classification and promotion. Its hold upon our more gifted children is shown by the fact that, although the distribution of intelligence follows the normal curve, there are approximately 10 times as many retarded pupils as accelerated ones in the schools. However much we may deprecate the disregard of individual differences which this system implies, no perfect method of classification and promotion has as yet been devised. The problem is still in the experimental stage.

The following study, which presents the results of a survey of the three first grades of the Horace Mann School, is concerned with this problem. Part I indicates the need for the survey, reports its general method and results, and suggests a plan for utilizing the data thus secured for purposes of reclassification. Part II¹ records the correlations obtained between the various measures employed in the survey, evaluates these measures, and gives a detailed account of statistical procedure.

PART I. THE NEED FOR THE SURVEY; ITS METHOD AND RESULTS

The initial composition of the three first grade groups.—Although in the provision of three first grades the Horace Mann School has administratively available the facilities for a relatively homogeneous classification of entering children, previous to the survey described in this article no carefully standardized procedure for this purpose had been developed.

¹Part II is to appear in the next issue of the *Journal*.

In general, children entering from the Teachers College kindergarten were placed in one of two classes, for convenience in this article designated as Class A and Class B, respectively, children of less maturity and ability to co-operate, as judged by the kindergarten teachers,² being placed in Class A, and those of more, in Class B. To the third first grade, designated as Class C, were assigned the children entering from the outside. Wishes of parents or accidental factors, such as a vacancy in one of the rooms, also might determine assignment. Moreover, Class C was designed to prepare a certain number of its pupils for third grade work. Hence, children of unusual maturity and ability, or of some previous school training, or children who had failed of promotion, were included in its constituency.

The unstandardized basis of selection thus resulted in considerable variation within a given class in chronological age, general intelligence, intellectual maturity, ability to make social adjustments, as well as in entirely unnecessary overlapping among the classes. Experimental instruction with the project method as a basis, and the giving of special attention to the development of desirable habits and attitudes, complicated the task of the teachers. Hence some means of reducing the heterogeneity of the individual classes seemed imperative, and a survey and comparative study of the three first grades was undertaken.

The data secured in the survey and their quantitative treatment.—In making the survey two kinds of data were assembled: (1) the results of psychological tests, both individual and group; (2) teachers' estimates.

The psychological tests comprised, as the individual test, the Stanford Revision of the Binet-Simon Tests, and, as the group tests, the Pressey Primer Scale and the Meyer Tests. The individual psychological examinations and the Primer Scale were given by the writers. Children who had previously been examined in the kindergarten by means of the Stanford Revision, as a rule, were not re-tested. The Meyer tests were administered by their author, Miss Helen Meyer.³ The teachers' estimates were rankings of their own

²In making these judgments the kindergarten teachers were doubtless influenced to a certain extent by the results of the Standard Revision of the Binet-Simon Tests, which had been given to many of the children.

³The Meyer Tests used in this survey should not be confused with the Myers Mental Measure constructed by Caroline E. Myers and Captain Garry C. Myers, also used for examining first grade children. For an account of the Meyer Tests, see an unpublished thesis, "Group Tests for Grades I and II," by Helen Meyer, on file in the Columbia University Library.

pupils in maturity and in ability in reading, the interpretation of the former trait being left to the teacher. There were thus available a maximum of five separate measures for each child: the Terman mental age, the score on each of the two group tests, and the teacher's ranking in two traits. These measures were subsequently incorporated into a composite score, each individual measure in this composite being given equal weight with the exception of the Terman mental age, which was doubled.⁴

The incorporation of such varied data in a single quantitative measure involved the reduction of all measures to a common basis. The one selected for this purpose was that of mental age. The scores in the various tests and the rankings of the teachers were converted accordingly. By the use of a more or less complicated statistical procedure⁵ the five resulting mental ages, weighted as previously indicated, were then added, and the average mental age found. This average mental age was the composite score finally employed.⁶

A comparative study of the three first grade groups.—The survey revealed, as had been anticipated, wide variation in the mental maturity of the children, their native intelligence, and their chronological ages, and a striking amount of overlapping among the three classes. The extent of this variation and of the overlapping will be apparent in Tables I, II, III, and IV.⁷

Table I presents the data relative to mental age. Here are given for each class the median mental age found by each of the measures employed, including the composite, and the lowest and the highest mental age, with the corresponding range covered in each instance.

Since mental age as indicated by the Stanford Revision is being

⁴These weightings were adopted after a consideration of coefficients of reliability for the different measures; they are reported in Part II.

⁵Described in Part II.

⁶When fewer than five measures were available in the case of individual children, the composite score was obtained from the smaller number, provided the available measures included the Terman mental age.

⁷In the first three tables the data relative to the variability and degree of overlapping are indicated merely by the lowest and the highest figure found for each item, the range between the lowest and the highest figure, and the medians. Table IV, however, presents by classes the total distribution of the individual children in respect to the most significant measures employed. For this reason it was thought unnecessary to compute the precise amount of variation and overlapping by refined methods, especially in view of the fact that however unsatisfactory range is from the statistical standpoint as a measure of variability, it is of very great practical significance for the purposes of our study. The "extremes" in our data imply acute maladjustment on the part of actual children, which, in view of the relatively little attention given by teachers to the problems resulting from individual differences, can scarcely be too much emphasized.

TABLE I.
DIVERGENCES IN MENTAL AGE.

	Median			Lowest Mental Age			Highest Mental Age			Range		
	A	B	C	A	B	C	A	B	C	A	B	C
Stanford Revision.....	7-4	7-11	7-5	5-11	6-1	6-6	9-8	9-3	9-3	3-9	3-2	2-9
										More than 4-5	More than 4-8	More than 3-11
Pressey Primer Scale....	7-6	7-6	8-4	6-1	5-10	6-7	Above 10-6 ^s	Above 10-6	Above 10-6		More than 4-5	More than 3-11
Meyer Tests.....	6-9	7-4	8-2	5-6	6-2	5-11	8-3	9-0*	9-0		More than 2-10	More than 3-1
Teacher's Ranking in Ma- turity	7-4.5	7-10.5	7-8	5-7.5	6-3	5-10	9-1.5	9-6	9-3	3-6	3-3	3-5
Teacher's Ranking in Abil- ity in Reading.....	7-4.5	7-10	7-10.5	6-2	6-6	6-6	9-0.5	9-0	9-3	2-10.5	2-6	2-9
Composite	7-5	7-11	7-10.5	6-3 (Rank 67)	6-5 (Rank 66)	6-8 (Rank 64)	8-10 (Rank 6.5)	9-3 (Rank 1)	9-2 (Rank 2)	2-7 (69.5)	2-10 (65)	2-6 (62)

^sThe highest mental age utilized in the table for converting scores in this test into mental ages.

considered in an increasing number of schools, the most important single factor for the placing of school children, the data in regard to this test are of especial significance. It is at once seen from a consideration of the medians and the extremes in ages for the three classes that the overlapping is almost total; moreover, that the average range for the three classes is approximately 39 months, and that in one class alone appear the extremes in mental age for the total group. That the great variability shown by the range between the extreme measures is characteristic of the group as a whole is confirmed by the standard deviations, which are for the three classes, 10 months, 8.5 months, and 10 months, respectively. Such variation in mental maturity is not more than could naturally be expected in a first grade; that it should be distributed apparently without reference to class groupings is the striking fact.

The evidence with reference to overlapping presented by the Stanford Revision is further supported by the data secured from the other tests and the teachers' rankings. The facts given by the composite, which takes into account all of the measures, are of most importance; they are also in essential agreement with those presented by the Stanford Revision, though the divergences found in mental age are less.

The rank numbers corresponding to the extremes of composite mental age in each of the rooms are especially convincing. A total of 67 children was ranked by means of this measure. Reference to the table shows that Class A included children ranking from 6.5 to 67; Class B, from 1 to 66; and Class C, from 2 to 64. Greater overlapping than these figures show is thus almost impossible.

The range in composite mental age for the entire group, regardless of grade, is three years, extending from six years, three months, to nine years, three months. The median mental age, this same measure being used, is seven years, nine months.

In Table II are given the data for intelligence quotients similar to the data already tabulated for mental age, although less complete, since intelligence quotients were calculated only on the basis of the mental ages derived from the Stanford Revision and from the composite. The variation in native intelligence of the children, as shown by the intelligence quotients, is more clearly indicative of mal-adjustment than the divergences in mental age. If the range in Terman intelligence quotients of the class which is the most

variable be considered, namely, 61, it is at once evident that a difference is found here which is as great as that between definite feeble-mindedness and very superior intelligence—a difference which is rendered none the less serious by reason of the fact that the variation falls in this instance between the limits of the lower ranges of normal intelligence and genius. The difference in the second group is scarcely less. Moreover, if the extremes in the composite intelligence quotients for each class, which are, perhaps, of even greater significance than the Terman, be considered (102 and 147 for Class A, 86 and 139 for Class B, and 94 and 130 for Class C) variations in intelligence in the three groups from normal intelligence to genius, from dullness to the upper ranges of very superior intelligence, and from the lower ranges of normal intelligence to very superior intelligence, respectively, are seen to be present.

TABLE II
DIVERGENCES IN INTELLIGENCE QUOTIENT

	Median Class			Lowest I. Q. Class			Highest I. Q. Class			Range Class		
	A	B	C	A	B	C	A	B	C	A	B	C
Stanford												
Revision....	130	119	116	95	86	94	156	143	137	61	57	43
Composite....	127	121	117	102	86	94	147	139	130	45	53	36

Table III presents for chronological age^a data of the same type as already presented for mental age and intelligence quotient. The differences indicated here are seen to be less than those found for mental age, and are naturally of less significance for the teacher. The overlapping is total for classes B and C, but less marked in the case of class A—a natural resultant from the custom explained in the earlier part of this article of placing in this class the younger children entering from the kindergarten of Teachers College. Although in general Class A has the youngest children, all three groups contain one or more children under six years of age, and one or more over seven years of age. The range in two out of the three groups is two years, and in the third, two years, one month.

TABLE III
DIVERGENCES IN CHRONOLOGICAL AGE

Median			Lowest Chronological Age			Highest Chronological Age			Range		
Class			Class			Class			Class		
A	B	C	A	B	C	A	B	C	A	B	C
6-0.5	6-8	6-11	5-2	5-9	5-9	7-3	7-9	7-9	2-1	2-0	2-0

^aCalculated to January 1st, as explained in Part II.

In Table IV is presented a distribution of the individual children by classes in mental age and intelligence quotient: the mental age employed is the composite, and the intelligence quotient, that derived from this mental age. The two dotted lines which divide the distribution into four parts indicate, respectively, the steps in which the median of the mental ages and that of the intelligence quotients lie.

TABLE IV

DISTRIBUTION OF MENTAL AGES DERIVED FROM COMPOSITE OF MEASURES, AND INTELLIGENCE QUOTIENTS CORRESPONDING

I. Q.	80-89			90-99			100-109			110-119			120-129			130-139			140-149		
Class	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
Mental Age																					
6-0 to 6-5							1	1		1			:								
6-6 to 6-11		1				2	1			1		1	1								
7-0 to 7-5						1	1	3	3	1	2	1	4								
7-6 to 7-11	1	1	1	2	1	1	5	1	4	1	1
8-0 to 8-5											2	2	:	4	2	1	1				
8-6 to 8-11												1	:	3	2		1	1	2		
9-0 to 9-5													:	1	1		1				

The Table indicates in detail the amount of variation and overlapping. An inspection of the grouping which results from the drawing of the two dotted lines reveals the fact that children from all three classes are represented in each of the following groups: (1) those below the median both in mental age and in intelligence quotient; (2) those in the median steps both in mental age and in intelligence quotient; and (3) those above the median both in mental age and in intelligence quotient. It is thus apparent that the procedure followed in the assignment of pupils to a given class has almost totally failed to take advantage of the opportunity afforded by three first grade groups in the school, for adaptation to individual differences in mental maturity and intelligence.

Naturally, we should expect to find approximately an equal number of children in each of the four divisions of the table indicated by the dotted lines; that is, about an equal number in the low mental age—low intelligence group, in the low mental age—high intelli-

gence group, in the high mental age—low intelligence group, and in the high mental age—high intelligence group. As a matter of fact, however, only two of these groups are clearly defined, namely, the low mental age—low intelligence group, and the high mental age—high intelligence group. Except for the five children falling in the high mental age—low intelligence group, all the children who may be presumed to belong in these two divisions fall in the steps in which the medians lie.

This at once raises the question as to whether some selective factor has not been operating to determine the distribution. In fact, it is highly probable that the other children whom we should have expected to find in the high mental age—low intelligence group, are already in the second grade. On account of their greater chronological age, they have entered school a year earlier, and have thus already been promoted more or less automatically; whereas the children of the same mental age, younger in chronological age, but of greater ability, although fully as able mentally to do the work of the second grade, still remain in the first. Conversely, the other children, whom we should have expected to find in the low mental age—high intelligence group, since they would be under six years of age chronologically,¹⁰ have not as yet entered school, although as able mentally to do the work of the first grade as the older, less able children already admitted. The distribution in Table IV thus affords evidence of the tendency to promote according to age rather than according to ability, and demonstrates the practical consequences of making chronological age so largely the basis of admission to school.

A Plan of Reclassification Utilizing the Data Secured

The answer to the question as to the best use to be made of the data gathered in the survey for purposes of reclassification is probably to be found in the facts presented in Table IV. Lines drawn at the end of the step 7-6 to 7-11 for mental age, and at the beginning of the step 120-129 for intelligence quotient, would suggest boundaries of three first grades, which would include children of relatively equal mental maturity and ability. It is thus possible to

¹⁰That is, under six years of age on January 1, the date to which the chronological ages of all the children were reckoned, as previously explained.

organize three groups fairly homogeneous from the standpoint of the data collected. In the actual carrying out of any program of reclassification, however, some adjustments would undoubtedly be necessary. Moreover, additional factors which have not been quantitatively measured in this survey, except in so far as they may to some extent have influenced the judgments of the teachers, should be taken into consideration in determining the placing of problematical cases. Among these factors may be mentioned chronological age, physiological development, health, nervous stability, and ability to make social adjustments.

Practical Consequences of the Survey

The complete returns of the survey were available for use at the close of the first semester. But on account of certain administrative difficulties, and the reluctance of the teachers to give up pupils whom they had come to know and understand, a reclassification on this basis was not attempted. Nevertheless, two consequences of considerable practical importance resulted: (1) the increased knowledge on the part of the teachers as to the nature and extent of the differences existing among their own pupils, and a correspondingly greater sympathy and insight into their individual needs; (2) the convincing demonstration of the necessity for gathering similar data for the children throughout the elementary school, looking toward a more scientific method of classification and promotion in the spring.

It is outside the limits of this article to give in detail the methods or the results of the survey of the entire elementary school carried on in the spring. A brief statement only must suffice. In the first place, the available records in psychological tests were assembled for all the pupils in the elementary school, including the children in the kindergarten who were to enter the first grade in the fall, and the chronological ages of the children for February 1st of the following year ascertained. These records were then utilized in computing the mental ages for that date.¹¹ In addition, teachers' rankings of the pupils as to their fitness for promotion were secured, and standard pedagogical tests in reading and arithmetic given to the children in grades III to VI, inclusive. Subsequently, the teachers'

¹¹The mental ages were computed to the mid-point of the following year, as described above, in order to make possible the formation of groups that would be more nearly homogeneous throughout the year for which the classifications were being made.

rankings were converted into mental ages¹² and the scores in the pedagogical tests into educational ages. After appropriate weightings had been assigned, these various measures were converted into a composite score for each pupil. This single figure then served as a comparative measure to be used as a basis for determining promotions.

The influence of this survey upon the constituency of the first grade groups for the following year will be of particular interest. The records of the kindergarten children were utilized in a tentative classification of first two grade classes differing noticeably in maturity and ability, to the third class being assigned the new children entering from outside the school. This method made possible any necessary adjustments as soon as the evidence from the testing of the new children entering in the fall and the judgments of the three first grade teachers could be taken into consideration in determining the permanent classification of the children for the year.

¹²According to the third method described in Part II.

RESULTS OF THE COMBINED MENTAL-EDUCATIONAL SURVEY TESTS

RUDOLPH PINTNER and HELEN MARSHALL,
Ohio State University.

In a previous article we have described the construction of the mental and educational survey tests and the significance of the ratings which can be derived from these tests.¹ In this article we give the results obtained by the use of these tests during the last two or three years and shall try to explain their value to the teacher and superintendent in helping to solve the practical problems of classification in the school.

Let us remind the reader that our index ratings run from 0 to 100, that 50 per cent. of the distribution lies between the index numbers 40-59, the upper and lower 23 per cent. between indices 60-79 and 20-39, respectively, the top 2 per cent. between 80-100 and the bottom 2 per cent. between 0-19. These index values are obtained readily from our tables given in the Manual of Directions.² The teacher has no computing to do after scoring the tests. After obtaining the total score, she looks in the proper column of the table and finds the index, and the index so found really interprets itself. Three valuable indices are obtained—(1) The Mental Index, showing the child's mental standing; (2) The Educational Index (age), showing the child's educational standing in comparison with other children of the same age; (3) The Educational Index (grade), showing the child's educational standing in comparison with other children of the same grade. The significance and value of these indices will become apparent from a discussion of some typical results in the 56 schools tested.

The Class Chart.

Figure I is a typical class chart, taken from the fourth grade of one of the schools. The chart is divided into five horizontal columns, which contain the five groupings of children according to the mental index: 0-19, Dull; 20-39, Backward; 40-59, Normal; 60-79, Bright;

¹Pintner, R. and Marshall, H. *A Combined Mental-educational Survey*. J. of Ed. Psych. Vol. XII, No. 1. January, 1921, pp. 32-43.

²Manual and Test blanks can be obtained from the College Book Store, Columbus, Ohio.

80-100, Very Bright. In the same way the chart is divided into five vertical columns, which contain the divisions according to educational index by age: 0-19, Very Poor; 20-39, Poor; 40-59, Average; 60-79, Good; 80-100, Very Good. The dots scattered on the chart represent the individual children in the class. Thus, for example, a child may have a mental index of 61 and an educational index of 51, and a dot is therefore placed at the intersection of these two lines. Interpreting this, we could say that the child is mentally bright, but is doing only average work educationally. In the same

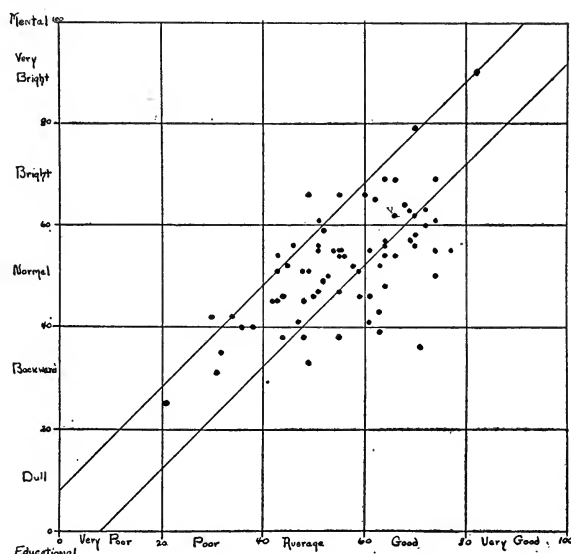


Figure I. Class chart showing mental and educational indices.

way a child whose mental index is 35 and whose educational index is 32 is backward mentally and is doing poor educational work for his age. In another case we find a child with a mental index of 39 and an educational index of 63. This child, although slightly backward, has succeeded in doing good educational work.

The diagonal lines on the charts enclose those cases whose index differences lie between minus 8 and plus 8, and, as explained in our previous article, include the middle 50 per cent. of the total distribution.³ These children may be considered as doing approximately

³Pintner, R. and Marshall, H. op, cit.

work which would be indicated by their mental ability. The upper triangle to the left includes those cases whose index differences are greater than minus 8. This means that all the cases in this triangle are doing less school work than their mentality seems to warrant. In this particular class there are seven cases lying in the upper triangle; that is, there are seven children who are doing decidedly poorer work educationally than they should be doing considering their mental ability. They represent wasted intelligence in the school. Cases of this sort should be scrutinized carefully by the

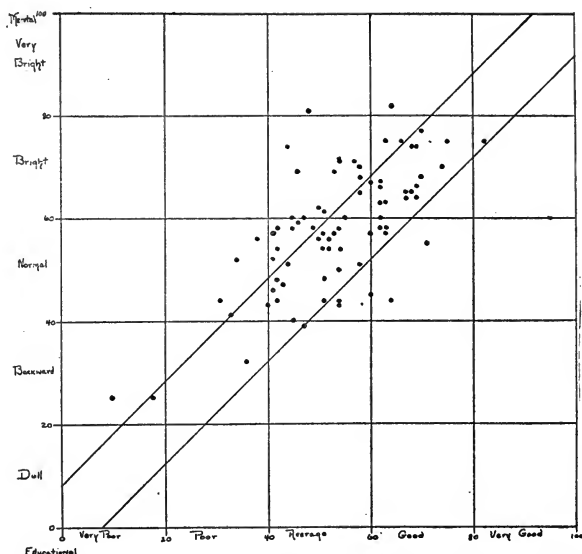


Figure II. Class chart showing mental and educational indices.

teacher from the standpoint of getting more out of those particular children than they are at present accomplishing. The corresponding lower triangle to the right includes those cases whose index differences are above plus 8. This indicates that all the cases in this triangle are accomplishing more than the average child of their mentality, this being probably due to good teaching, a good school, a good home, regular attendance, placement in proper grade, and the like. In this particular class we note there are 22 such cases and, comparing this number with the seven cases in the upper triangle, we should conclude that the class as a whole is accomplishing edu-

cationally more than is usually accomplished with children of their mentality.

In spite of the general good showing of the class under discussion, there are other possible suggestions that we may draw from the chart. In the four upper right-hand squares of the chart we note that there are 12 children who are all above 60 both mentally and educationally. They are all bright or better mentally and are all doing good or better educational work for their age. If this educational work is also rated as good educational work for the grade they are in, they should certainly be considered for immediate promotion. As we shall see later, we must take into consideration their

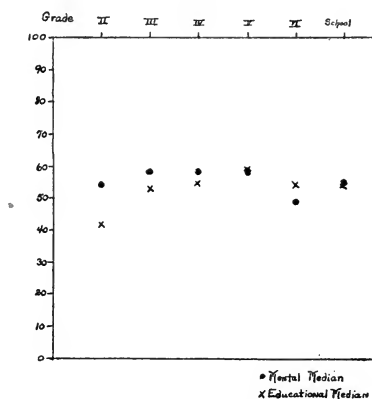


Figure III. School chart showing median indices for each grade.

educational indices according to grade and in accordance with these indices recommend an extra promotion.

Figure II shows another class chart. This time we note a large number of cases in the upper left-hand triangle. There is a much greater waste of intelligence in this class than in the previous one. Although the class as a whole shows a large majority of its pupils possessing normal mentality or above, nevertheless their educational attainment is not commensurate with their mental ability. The whole class lags behind what it should accomplish, and there are some striking individual examples of waste. At the top of the chart we note two children classed as very bright by the mental tests. One of these is only accomplishing average educational work and the other good educational work. Both of these should be in do-

ing very good educational work. Obviously there is great need for investigation here. And the same is true of many bright and many normal children. Is the teaching of this class poor, or are the standards set by the teacher or school too low? Something evidently is wrong and our tests have brought it to light very distinctly.

School Chart.

Figure III summarizes the work of an entire school. Thus the median mental index for the second grade is 54 and the educational median is 42; that is, the class as a whole is slightly above the average in mentality, but is doing poorer educational work than is to be expected. An index difference of minus 12 is undesirable, and the grade needs special attention in order to bring its educational work up to standard. In this particular instance the discrepancy is probably due to the crowded conditions in the school. The class was seated in a basement room, with inadequate lighting. Moreover, the children attended school only half a day. Grade IV, on the other hand, has a mental median of 58 and an educational median of 55. The difference here of minus 3 is so small as to be negligible, and both medians are well above average.

The chart also shows the medians for the whole school. As is to be expected, the difference is less than that of any specific class, since it combines the work of so many classes. The mental median for the school is 55, and the educational median is 54. That means that the school is somewhat above the average both mentally and educationally, and the difference of minus 1 is negligible.

The significance of a chart of this kind will be obvious to any principal. It shows immediately the grades which need more attention and investigation, both as to quality of teaching and proper classification of pupils. Wherever the educational index is below the mental to the extent of 3 or 4 points, there is obvious need for further investigation. In this particular school grades II and III are not properly adjusted. There is too large a discrepancy between their mental and educational indices.

Comparison of Schools.

It has been the practice in the past to rate schools primarily on the basis of gross accomplishment. It is undoubtedly desirable to know that the pupils in a given school are accomplishing so much

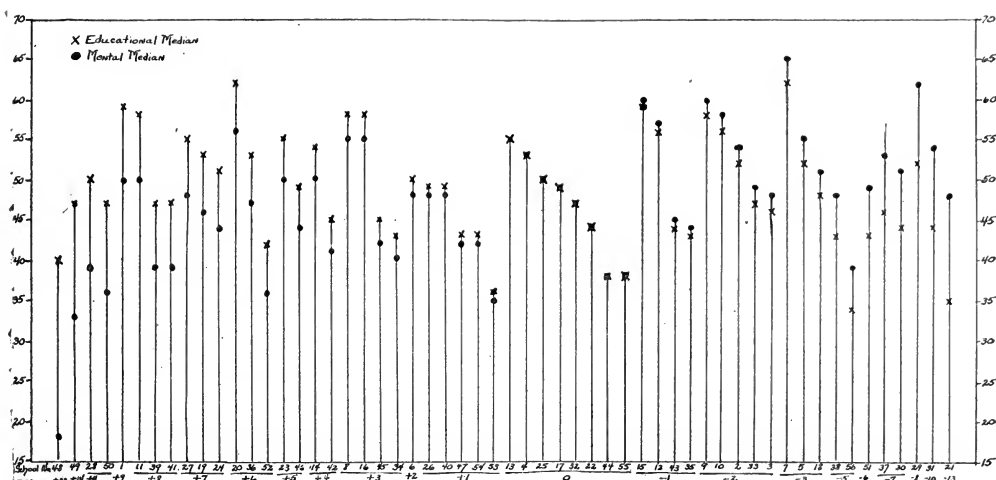


Figure IV. Comparison of schools.

more or less than the pupils in another school. And, all other things being equal, this would be a perfectly satisfactory method of evaluating schools. Such a method, however, neglects one very important factor in the equation, namely, the pupil-material with which various schools have to deal. School A may be doing exactly the same grade of work, educationally, as School B. Does that mean that the two schools deserve equal credit? School A is a small rural school, with an apparently selected group of children, whose median mental index is 60. School B is a larger school, whose three teachers must divide the work of handling eight grades. Moreover, the median mental index for School B is only 51. Instead of commending School A for its educational work, we will come to expect even better work from it, work commensurate with its pupil-material.

In preparing Figure IV, the schools were first ranked on the basis of gross educational accomplishment. Thus—

Schools numbered.	Median educational indices (grade).
1-4	60 and above
5-14	55 to 59
15-23	50 to 54
24-39	45 to 49
40-48	40 to 44
49-56	35 to 39

The numbers of the schools rank them from 1 to 56 according to their educational (grade) index; that is, according to the educational accomplishment of the children when compared with children of like grade.

The schools were then arranged according to the difference between the Median Educational Index (age) and Median Mental Index. The median educational index as shown on the chart is, therefore, a rating which compares each child with children of his own age, whereas the rank numbers of the schools show the gross educational attainment regardless of age. As in rating individuals, the school with the highest positive index difference receives the greatest amount of credit, without regard to any other measure. An examination of the table and a study of the relative position of the schools in comparison with their absolute rank will yield some interesting information. For example, school No. 1, which ranks first educationally, is also a good school on this basis, since it has an index difference of plus 9. On the other hand, school No. 52, which ranks low on the basis of absolute gross accomplishment, nevertheless shows a difference of plus 6, and deserves credit for having accomplished so much more than its median mental index would promise. School No. 48 is an interesting case. This is a small rural school, with one family of low mentality contributing nearly half of the pupils. As a result, the median mental index for the school is only 18 (Dull). In spite of this evident handicap, the teacher has succeeded in raising the educational index median to 40.

In contrast to these schools, Nos. 2 and 3 are good examples of schools which, while doing very acceptable work educationally, yet fall short of the type of accomplishment which could be demanded of the pupil-material. It is safe to assert that with increased effort on the part of teacher and pupils, with improved home or school conditions, these schools could raise their median educational indices.

Again the chart is valuable in showing the great differences in mental ability and in educational attainment in different schools. Looking at the dots on the chart we note how great the difference in mental ability is. Schools No. 7 and 29 have the highest median mental indices, yet both of these schools are failing to take adequate advantage of the mental ability of their pupils. There are many schools with higher educational indices than No. 29 which do not

possess as good mental material. This type of chart can be strongly recommended to the school superintendent for a study of the schools under his direction.

Practical Problems to Be Solved by the Tests.

On the basis of the results of these tests, it is possible to make definite recommendations to schools for the purpose of increasing educational efficiency. Children whose index difference is minus 9 or greater are obviously not working up to their mental ability. Many bright children will be found in this group, as can be seen from a study of the upper left-hand triangles of Figures II and III. The percentage of children in this group represents the amount of waste in the school. It would be necessary then to investigate each case thoroughly and make every effort to remove the cause of lowered working power. The causes probably operating in such conditions are:

- (1) Misplacement in grade. The child may be too far advanced, and therefore baffled by the amount of strange material presented to him. Or, as is more common, he may not be far enough advanced, in which case he will find the work too easy and develop habits of laziness which it will later be difficult to destroy.
- (2) Lack of interest, laziness and similar emotional and temperamental factors.
- (3) Bad physical conditions.
- (4) Bad or undesirable home conditions.
- (5) Crowded conditions in the school.
- (6) Poor teaching, which fails to arouse the best that is in the child; dislike of the teacher.

While all of these conditions are such that they may be bettered by the school, it is the problem of grade misplacement which comes most definitely within the scope of these tests. In the case of children whose educational (grade) index is 70 or above, if at the same time the mental index is 70 or above, immediate promotion is desirable. Such children are simply forming habits of indolence if they remain in their present grade, and are fully capable of handling the work of the next grade, since in every case an index of 70 in one grade indicates a score equivalent to an index of 50 or more in the next higher grade. Therefore, promote immediately to the next

grade all children with both educational (grade) and mental indices of 70 and above.

In addition to this immediate promotion, those children having educational (grade) indices between 60 and 69 and mental indices of 60 to 69 should be coached for an extra (or double) promotion the following semester. They will undoubtedly be capable of doing the work, since an index of 60 in one grade is equivalent to an index of 40 or above in the grade next higher.

In addition to coaching these children for prospective promotion, every child with an index difference of minus 9 or below, not included in the two preceding groups, should receive special attention from the coaching teacher, not with reference to promotion, but merely to make them accomplish work worthy of their mental ability. These children will range all the way from Very Bright to Very Dull mentally. In determining the special subjects in which these children are deficient, it will be profitable to supplement the teacher's judgment by reference to the various divisions of the Educational Test. It must be kept in mind that reading ability is fundamentally necessary in all parts of this test.

A special class of bright children should be formed when there are a number of such children in about the same grade. This class would probably include all children with mental indices of 60 or above, whose educational (grade) index does not fall below 50. If such a group is kept as a separate unit, it can be advanced through the school system at a more rapid rate and with profit to them, since they will not be held back by the slower members of the class. On the other hand, the children left in the original class will not be competing with the high standards of this group, and will be able to accomplish better work as a result of the separation.

In the same way a slow-moving unit of dull children should be formed wherever conditions demand it. Children with a mental index of 39 or below are seldom able to keep up the regular pace of school work. It is this group which furnishes the bulk of our retarded children, who subsequently become dissatisfied with school and prove disturbing elements both in school and outside. If we recognize frankly that these children should be allowed to move more slowly, we will go far to better conditions and increase their happiness. Wherever it is possible considerable handwork should be taught in these slow-moving classes. Often a backward or dull child

can be interested in the more academic subject by means of a judicious mixture of motor activities.

Conclusions.

Mental or educational tests alone are not adequate for a thoroughgoing survey of a school or school system. Such tests show wide differences in the educational attainment and in the mental ability of the pupil-material. A real diagnosis of the difficulties existing in any particular instance can only be arrived at by a combined mental-educational survey. The evaluation of the educational findings in the light of the mental findings is the only adequate and scientific method of procedure. Such an evaluation will show where the real discrepancies exist. We shall find waste in pupils of all degrees of ability and, indeed, more often in those of superior ability. We shall find that many schools whose educational level seems passable or good are really inefficient and wasteful of the splendid pupil-material they possess. Combined mental-educational surveys will help to correct these deficiencies and to lay the emphasis where it really belongs.

MASSED VS. DISTRIBUTED EFFORT IN LEARNING*

L. A. PECHSTEIN,
University of Rochester.

Ebbinghaus, Yost, Browning, Brown and Washburn, Murphy, Ulrich, Pechstein, Lashley, Carr and Cummins have contributed experimentally to the question whether it is more efficient to learn a problem with the learning effort being massed, i. e., continuous in time, or broken by periods of changed activity. Everyone contributing to the discussion agrees, in general, that distributed practice is more efficient than massed, the tendency being to generalize for various types of learning material, such as motor, nonsensical and meaningful. Specialized discussions by Carr (1) and Cummins (2) raise the question whether the efficiency of distributed effort is confined to certain stages in the learning process or whether this mode of acquisition is uniformly effective for all stages in the development of a habit, they answering that distributed effort is of greater efficiency during the early stages. Lashley, Colvin and all educational psychologists comment upon the neurological aspects of distributed learning, although Lashley (3) seems to take exception to the doctrine of a gradual "setting" of the nervous connections between practice periods.

The purpose of this paper is to show that the entire question is tied up with a second, namely,—is the learning material mastered as a whole or in parts?

The experimentation is restricted to the field of motor learning, the learning problem being a very difficult maze reported three years ago (4). This maze subdivides into four sections highly comparable in basic features and difficulty, these being so arranged that the maze can be learned as a whole, in parts, or in any combination of parts. The learning criteria are the number of trials required to secure four out of five perfect runs, learning time and learning errors, the latter being those conventionally recorded in maze experimentation. The experimental technique is exactly the same as reported in earlier articles.

Earlier experimentation with this particular maze naturally had shown that human adults learned the problem far more economically

*Read at the annual meeting of the American Psychological Association, December 28-30, 1920.

when their effort was distributed at the rate of one trial per day rather than massed (5). In fact, the maze had proved too difficult for certain human subjects to learn under massed conditions. Also, it was shown to be more economical for these adults to learn their problem as a whole rather than to learn the four units separately, finally connecting them in serial order, provided the learning effort was distributed in each case. Because of the difficulty of the maze and the way in which the rat is organized, it was naturally impossible to secure massed learning results from the rats. The initial step of the present experiment was to shorten the maze problem, hereby to discover whether rats could be taught a simple problem under massed requirements; subsequently, to study human learning under comparable conditions.

Because the particular maze used in my whole-part experimentation is made up of four short and separate maze units, these simple units were employed to test the capacity of the rat to learn under massed conditions. After being handled and fed in the food box for a standardized period of 10 days, feeding was skipped for 48 hours, whereupon the training was begun. Each rat learned the first short maze section with apparent ease, although the amount of learning effort expended far surpassed what I had supposed the white rat capable of continuously directing. A short feeding was allowed after all in the group had mastered the first section. On the following day Section II was learned with ease under massed conditions, this being followed by the mastery of Sections III and IV upon the third and fourth day, respectively.

Upon the day following the learning of the last section, each rat was set the problem of connecting the units into proper serial order. Not only did each member of the group effect the perfect connections under massed conditions, but several made the complex step without securing a single error. Herein is the most striking feature of the entire situation. When humans or rats are taught these short units under distributed conditions, the complex act of connection proves extremely difficult. *When these units are learned under massed conditions, not only is this first learning task easy, but the hard act of connection becomes extremely simple.*

Comparable results held when human adults were tested under like temporal conditions.

An inspection of Table I points the way to several significant conclusions:

TABLE I.—A table to show the mean number of trials, time and errors of two groups of rats (nine per group) and two groups of humans (six per group) in learning Maze A by the "part" method. The records for distributed learning appear first in each section, the massed learning records second. The third item in the total is for groups learning the Maze as a whole under distributed conditions, the last as a whole under massed conditions. For estimating total runs, each sectional run is counted as one-fourth the entire (I-IV) run.

Section.	Rats.			Humans.		
	Trials.	Time.	Errors.	Trials.	Time.	Errors.
I	34	470"	52	6	198"	24
	11	256"	27	5	93"	12
II	2	33"	3	3	47"	10
	5	51"	6	1	45"	9
III	14	127"	14	2	49"	5
	7	72"	12	2	28"	4
IV	9	111"	11	1	25"	7
	1	11"	2	2	28"	5
I-IV	15	1166"	119	20	901"	191
	4	439"	88	8	344"	78
	30	1907"	199	23	1220"	237
	10	829"	135	10	538"	108
	27	4174"	217	12	641"	126
Total.....	30	1250"	260

First, provided the maze problem is short, it is more economical to mass learning effort than to distribute it, irrespective whether economy is estimated in terms of trials, time or errors.

Second, for subsequently learned short runs (allowing transfer possibilities), massed effort continues, in general, to be preferable, it hereby appearing that, if problems are thoroughly learned before subsequent ones are undertaken, the transfer is positive rather than negative.

Third, in connecting short maze patterns learned as separate units, the complex act of connection is not only possible in a massed program, but is accomplished with very great economy, just so long as the units have been learned as massed effort problems.

Fourth, the longer and more difficult the problem, the more advisable to break it up into units and learn both the units and the connection of these under massed conditions, it being uneconomical to learn the hard problem as a whole, irrespective whether effort is massed or distributed.

Fifth, it is clear that the question of massed vs. distributed learning is tied up with the question whether the difficult problem is to be learned as a whole or in parts. The hard problem becomes easy if it is learned under massed conditions by the part method and in no other way; it remains hard if it is learned as a whole under massed or distributed conditions, or even in parts under distributed conditions.

Sixth, these results hold for motor learning of the maze type, both for selected animals and the human adult.

The explanation of the conclusions is not far to seek. It is not necessary to comment upon the advantages inherent in any part method, such as the full utilization of transfer possibilities, the diminishing returns secured as the problem is lengthened, etc., since these are, presumably, fairly constant under both massed and distributed learning. It is essential to show why an easy problem is best learned under massed conditions and why several so learned can readily be united through massed effort. Explanation rests here, I believe, upon two principles operative in all learning.

The first of these is the principle of *elimination*. "It connotes the detection of critical points in the problem, the careful study of all the details, the formation of proper associations, the rejection of others, etc. Consciousness is here at white-heat. The longer and more difficult the problem, the greater the task upon the learner to see the many details of the problem and to learn to eliminate the possible faulty reactions. Confusion, hesitation, emotional conditions all operate to delay the learning" (6). If the problem is short and easy, the principle of elimination operates with maximum effectiveness. No emotional complex is aroused to disturb the organism in utilizing all his problem-solving, adjustmental powers; the relative fewness of possible errors makes these lie within the organism's powers of mastery; the learner has available the energy required for the successive explorative trials, and this is utilized; if given one trial only, the learner is far within the limit of his available strength. Massed effort upon a long and hard problem strains the power of the organism and elimination ceases to operate efficiently; massed effort upon an easy problem challenges the organism to work within normal and proper limits, and elimination operates efficiently; distributed effort upon an easy problem allows the eliminative principle to operate, but does not secure that total

learning efficiency normally expected after the preliminary warming-up stage; distributed effort upon the hard problem gives opportunity for the eliminative principle to operate, but not always at its best, since the difficulty of the problem generally strains the organism beyond the normal limits of its power.

The second learning principle involved is that of *mechanization*. "This final stage of learning is no longer explorative and eliminative, but rather, mechanizing and rendering habitual the entire activity. Whole method learning presents so many critical details that the principle of mechanization is not only delayed in being given an opportunity to operate, but is repeatedly nullified by the re-injection of the highly conscious eliminative principle" (6). This condition maintains both for massed or distributed attack upon the hard problem. The part method utilizes the two methods to best advantage. The explorative or eliminative principle is operative when the details of the short and easy parts are being grasped, it being made clear that the best elimination is secured by massing the learning of each short part. Then the time becomes ripe, logically and psychologically, for the mechanizing principle to operate. It secures the mechanization of each unit and their connection, herein being concerned only with rapidly welding the several unit habits. Psychologically speaking, the runner who hesitates in this stage is lost. By rapid and consecutive runs he forced the required union. Fortunately, if he has learned the units as massed problems, his act of connection is very easy, while it is extremely difficult if the unit habits are set up under distributed conditions. Why? Because in working steadily through a massed period upon short problems he has come to set up habits of long application, exactly what is required for the complex act of connection. The distributing learner, only having one trial per day upon short units, has acquired habits of short application, these almost rendering impossible the final connection and mechanization of the units into a perfect total habit.

The whole method never secures complete utilization of the principles of elimination and mechanization, the loss being greater for massed learning. The part method secures full utilization of the principles, provided the learner masses his effort.

Part method and massed learning—so long anathema to the pure and educational psychologist—may not be understood one apart

from the other; employed together, they make the best arrangement for learning difficult motor problems.

BIBLIOGRAPHY.

- 1.—Carr, Harvey. Distribution of Effort. *Psych. Bull.*, 16, 1919, 26.
- 2.—Cummins, R. A. Improvement and the Distribution of Practice. Teachers College, Columbia University, Contribution to Education. No 97 (1919).
- 3.—Lashley, K. S. A Simple Maze: With Data on the Relation of the Distribution of Practice to the Rate of Learning. *Psychobiology*, I, 5, 1918, 353.
- 4.—Pechstein, L. A. Alleged Elements of Waste in Learning a Motor Problem by the Part Method. *Jr. Ed. Psychol.*, 1917, VIII, 303.
- 5.—Pechstein, L. A. Whole vs. Part Methods in Motor Learning. A Comparative Study. *Psychol. Rev. Mon. Supp.*, Vol. XXIII, No. 2, pp. 59 sq.
- 6.—Pechstein, L. A. Whole vs. Part Methods in Learning Nonsensical Syllables. *Jr. Ed. Psychol.*, 1918, IX, 387.

AN EXPERIMENTAL STUDY OF THE VALUE OF WORD STUDY

V. A. C. HENMON,
University of Wisconsin.

During the first semester of 1919-20 there was conducted in the Madison High School, with a part of the sophomore class, an experiment in word study as a substitute for the regular work in English composition and literature. A group of 54 pupils, designated in this paper as the word study group, did intensive work in word study and analysis under the direction of Miss Leslie Spence. The work was based on a syllabus prepared by Miss Spence, and was carried on for approximately 12 weeks of the semester. The remainder of the semester was given to literature and composition.

Mr. Volney G. Barnes, principal of the school, while confident from personal observation and judgment that the course was valuable, raised the question with the writer as to whether it was not possible to apply tests which would demonstrate objectively the values that might have accrued from the course and prove or disprove the correctness of his opinion as to the desirability of its continuance. This was in the nature of a challenge to experimental education and was gladly accepted. The following study, made with the assistance of Miss Jane H. Butt, is the result.

The problem presented was to determine as accurately and definitely as possible the specific outcomes which such a word study course might be expected to realize, and then to select or construct tests which would measure these outcomes as adequately as possible. Moreover, for comparison, a group of pupils must be selected of equal ability in general scholarship and with equal amounts of foreign language who have had the regular work in composition and literature. The tests should, if possible, be of such a nature that they could fairly be given to this non-word study group also.

What, then, are the specific outcomes that might be expected of an intensive word study course? We may expect it to function in at least four specific ways:

First—In increase in vocabulary.

Second—In increase in ability to give meanings accurately.

Third—In increase in ability to choose words discriminately.

Fourth—In increase in ability to read difficult prose understandingly.

These would doubtless be accepted as the major outcomes of the course whatever other values might be claimed.

The tests employed to measure them were:

1. Terman's Vocabulary Test, 100 words, given as a group test with these instructions: "Show that you know what each word means either by giving a definition of it, a synonym of it, or by using it in a sentence in such a way as to show that you know its meaning."
2. Thorndike's Visual Vocabulary, scored both according to Thorndike's method and the total number right.
3. A special test of 25 words selected from a list of 100 words where a knowledge of roots, prefixes and suffixes would have an opportunity to function, but which should also be words with which pupils of sophomore high school grade might reasonably be expected to have come in contact.
4. Trabue's Completion Scale L, used to test ability to discriminate in the choice of words.
5. Tests 1a and 1b of the Thorndike Intelligence Examination, Part III, used to measure ability to read rather difficult prose understandingly.

These tests were given to all members of the sophomore class, approximately 350. Each pupil of the word study group was paired with another of equal scholarship based on freshman marks and marks of the first semester of the sophomore year and of equal amount of language work in Latin or French. There are, then, two groups, the Word Study and the Non-word Study, approximately equal in general scholarship and in school marks in English, Latin and French.

Table I gives the essential facts for comparing the groups both in scholarship and in the tests. Whatever advantage there may be in the two groups in scholarship, except in Latin, is in favor of the non-word study group as indicated by the differences in averages. There were, however, only six pupils in each group who had had Latin. It will be noted that the differences in scholarship are all well within the probable error of the difference except in the case of the English marks, which definitely are in favor of the non-word study group.

The results in the tests show in each case a positive difference in favor of the word study group. That the differences, while small, are significant is indicated by the fact that they are in all cases larger than the probable error of the difference, four times the probable error or more in four of the tests. In the vocabulary tests, where we might look for the greatest showing, we find the greatest differences. This is particularly true in the word meaning test, as might be expected. The reading and language ability tests, Thorndike 1a and 1b, and Trabue Completion, however, give differences that are substantial and significant unless one sets a standard of difference to be significant higher than four times the probable error.

The specific question proposed at the outset of this investigation was this: "Is a special course in word study for high school sophomores worth while and should it be confined?" The results, if taken at their face value, would indicate an affirmative answer. Before drawing such a conclusion, however, certain questions arise. Was the statement of the outcomes adequate or are there other important outcomes that were overlooked? (1) The course might be claimed to possess disciplinary value quite apart from the specific outcomes enumerated. In the light of present-day educational opinion this would, in any case, be a doubtful basis for the introduction or continuance of such a course. (2) It might serve as an introduction to the study of foreign language or arouse an interest in it. This effect might be shown in elections next year and in progress in foreign language work. (3) It might develop that rather intangible thing called language consciousness. (4) It might function in study habits and, in particular, give training in use of the library. No attempt has been made to evaluate these possible results experimentally.

Are the tests adequate and valid? They do measure very definite outcomes, but they do not, of course, measure what the word study group may have lost in literary appreciation or improvement in composition which may have come from the regular work in literature and composition. It should be borne in mind, however, that the specific outcomes of word study are, at the same time, the major objectives in composition and literature. The enlargement of vocabulary, increase in discriminativeness in the use of words, and ability to read understandingly are surely the major purposes of English study of any sort. While more specific tests might have

been devised for the word study group, the necessity of tests that could be given to the non-word study group fairly led to the selection so far as possible of tests already standardized. While it would have been desirable to give a general intelligence test in order to insure equality in the groups, the four sets of scholarship ratings were deemed sufficient evidence.

Even so, the question proposed is difficult to answer on a basis of these results, or, for that matter, on the basis of any study that might be made. The answer depends upon one's educational philosophy as to the aims and purposes of any sort of education and training. If, however, we grant the major propositions that the enlargement of vocabulary, increased ability to discriminate in the use of words, and finally increased ability to read understandingly are themselves definitely worth while and worth making definite efforts to secure, then the question would be answered in the affirmative. Genuinely significant differences in these abilities have been shown to result from less than a semester of word study.

GROUP TESTS OF INTELLIGENCE: AN ANNOTATED LIST

J. CARLETON BELL,

Maxwell Training School for Teachers, Brooklyn, N. Y.

Since Alfred Binet published his "Measuring Scale of Intelligence" in 1908 the interest in tests of intelligence has grown apace. For the first few years after the Binet Scale became known in this country attention was directed chiefly to the adaptation and revision of it to fit American conditions. In this period we have the revisions of Goddard, Huey, Wallin, Kuhlmann and (most recent and best known) Terman. As a derivative of these attempts at revision we also have the "Yerkes Point Scale for Measuring Mental Ability." The Binet Scale, however, and all of its revisions demand the individual examination of the child, and to make individual studies of several hundred or several thousand children requires such an expenditure of time and energy as practically to prohibit the incorporation of mental testing into the regular routine of school work.

Out of this situation arose the demand for Group Tests of Intelligence—a demand greatly intensified by our sudden entry into the World War, and by our need of the help of psychology to 'put the right man in the right place.' The wide publicity given to the wholesale testing of our recruits acted as a tremendous stimulus to the development of Group Tests. As a result the past two years have seen the publication of an almost bewildering array of tests, and there is now great need for comparison, analysis and evaluation. In the following the writer has attempted merely to list the tests of which he knows, and to comment briefly on the more important of them. It is likely that the list here compiled is far from complete:

1. Mrs. Sidney L. Pressey. Mental Survey Tests, Primer Scale. Department of Psychology, University of Indiana, Bloomington, Ind. A group of four tests (dot pattern, classification, picture form board and picture absurdities) designed for grades 1-3. The tests involve neither reading nor writing, enlist the active co-operation of the children, and give a correlation of .58 with teachers' estimates.
2. Arthur S. Otis. Otis Group Intelligence Scale; Primary Examination. World Book Company, Yonkers, N. Y. Eight picture tests (directions, substitution, missing parts, maze, sequence, simi-

larities, etc.) for grades 1-4. Issued in two equivalent forms, A and B.

3. M. E. Haggerty. Intelligence Examination. Delta 1. World Book Company, Yonkers, N. Y. Six tests, with a fore-exercise for each one. The tests are directions, copying designs, picture completion, picture comparison, symbol-digit, and word comparison, and are designed for grades 1-3. They were used in the Virginia School Survey, and give a correlation of .68 with teachers' estimates.

4. Kingsbury Primary Group Intelligence Test. Bureau of Educational Research, University of Illinois, Urbana, Ill.

5. Walter F. Dearborn. Group Intelligence Tests for Primary Grades. J. B. Lippincott Company, Philadelphia, Pa. Sub-title, "Games and Picture Puzzles."

6. Frances Lowell. A Group Intelligence Scale for Primary Grades. *Journal of Applied Psychology*, Vol. III, September, 1919, 215-247. Twenty-five tests arranged in groups of five for each year for the chronological ages 5 to 9, inclusive. The tests are taken largely from Binet, Kuhlmann, and other standard sources, and adapted to use with groups.

7. Caroline E. and Garry C. Myers. The Myers Mental Measure. The Sentinel, Carlisle, Pa. (See *School and Society*, September 20, 1919, 355-360.) For grades 1-8. An outgrowth of experience with the Army Mental Tests. Requires only 20 minutes to give, and is said to show a correlation of .80 with the Standard Revision of the Binet Scale.

8. Rudolf Pintner. The Mental Survey. New York: D. Appleton & Co., 1918. Pp. 116. Six mental tests, taken chiefly from Whipple, and standardized as group tests on over 3000 school children of chronological ages 6-16. Elaborate tables of percentile norms are given.

9. Grace Arthur and Herbert Woodrow. An Absolute Intelligence Scale. *Journal of Applied Psychology*, Vol. III, June, 1919, 118-137. A battery of nine tests (memory span, easy opposites, hard opposites, substitution, word building, language completion, anagrams, cancellation, and comprehension) standardized by elaborate statistical treatment to give norms in absolute points for the chronological ages 6.5 to 13.5.

10. Rudolf Pintner. A Non-language Group Intelligence Test. *Journal of Applied Psychology*, Vol. III, September, 1919, 199-214.

A series of six tests (imitation, easy learning, hard learning, drawing completion, reversed drawings, and picture reconstruction) standardized on the basis of point scores for the chronological ages 7 to 13.

11. Edward L. Thorndike. A Standardized Group Examination of Intelligence Independent of Language. *Journal of Applied Psychology*, Vol. III, March, 1919, 13-32. Eight tests (digit-symbol, lines dividing surfaces, picture completion, forms completion, picture analogies, spatial relations, memory of objects, and easy computation) derived from the Beta Army Tests, and applied to groups ranging from feeble-minded (mental age 7.5) to superior adults.

12. The Illinois Examination. (See B. R. Buckingham and Walter S. Monroe, A Testing Program for Elementary Schools, *Journal of Educational Research*, Vol. II, September, 1920, 521-532.) A series of tests including analogies, arithmetical problems, sentence vocabulary, substitution, verbal ingenuity, arithmetical ingenuity, and synonym-antonym, for grades 3 to 8.

13. William Henry Pyle. A Manual for the Mental and Physical Examination of School Children. *University of Missouri Bulletin*, Vol. 21, No. 12, February, 1920. The mental examination consists of eight tests (logical memory, rote memory, substitution, free association, opposites, word building, completion, and analogues) with norms for the chronological ages 8 to 18.

14. S. L. Pressey and L. W. Pressey. A Group Point Scale for Measuring General Intelligence, with First Results from 1100 School Children. *Journal of Applied Psychology*, Vol. II, September, 1918, 250-269. The scale contains 10 tests, as follows: Rote memory, logical selection, arithmetic, opposites, logical memory, word completion, moral classification, dissected sentences, practical information, and analogies. There are tables of norms for chronological ages 8 to 17.

15. M. E. Haggerty. Intelligence Examination, Delta 2. World Book Company, Yonkers, N. Y. An adaptation of the Army Intelligence Examinations for school children of grades 3 to 9, for use in the Virginia School Survey. The examination consists of six tests (sentence reading, arithmetical problems, picture completion, synonym-antonym, practical judgment, and information), with age and grade norms in terms of total scores for ages 8 to 15.

16. National Intelligence Tests. Scale A (arithmetical reason-

ing, sentence completion, logical selection, synonym-antonym, and symbol-digit) and Scale B (computation, information, vocabulary, analogies, and comparison). World Book Company, Yonkers, N. Y. These scales are designed for use in grades 3 to 8. They are the outgrowth of the Army Intelligence Tests, and were prepared under the auspices of the National Research Council by a committee of psychologists consisting of M. E. Haggerty, L. M. Terman, E. L. Thorndike, G. M. Whipple and R. M. Yerkes, chairman. The General Education Board appropriated \$25,000 for the standardization of these tests, and they are probably the most reliable of any group tests of intelligence. Each scale is furnished in five forms of equivalent difficulty.

17. Paul R. Stevenson. Omaha Group Test of Intelligence. Bureau of Research, University of Omaha. Eight tests, containing from five to fifteen exercises each. They include likeness and difference, correction of statement, arithmetical problems, disarranged sentences, following directions, synonyms-antonyms, analogies, and range of information. They are intended for grades 3 to 9, and require 20 minutes to give. In the printed directions one or more crucial words are omitted and must be written in at the dictation of the examiner. This is designed to prevent pupils working ahead.

18. Guy M. Whipple. Group Test for Grammar Grades. Public School Publishing Company, Bloomington, Ill. These tests (arithmetic, completion, substitution, reasoning, punched-hole test, and proverbs) have been tentatively standardized for grades 4 to 7, but difficulties of administration make their use somewhat unsatisfactory.

19. Walter F. Dearborn. The Dearborn Group Tests of Intelligence, Series II, General Examinations, 4 and 5, for grades 4 to 9. J. B. Lippincott Company, Philadelphia. The title printed in large type at the top of the first page is "Games and Puzzles." There are ten tests, three of them verbal. They are: Sequence of relations, order of terms, picture form board, completion, direction, pictorial representation of terms, mazes, proverbs, picture absurdities, and computation.

20. Sidney L. Pressey. Mental Survey Scales. "Cross-out Tests." A Brief Group Scale for Measuring General Intelligence. Department of Psychology, Indiana University, Bloomington, Ind. Four simple tests (verbal ingenuity, logical judgment, arithmetical

ingenuity, and moral judgment) so arranged that the pupil shows that he has comprehended each part by striking out a single superfluous word. The author has recently issued norms for grades Low 4 to High 8.

21. Arthur S. Otis. *Otis Group Intelligence Scale; Advanced Examination*. World Book Company, Yonkers, N. Y. The oldest and perhaps the most widely used of the group tests of intelligence. The series consists of 10 tests, each containing from 20 to 30 tasks. (See *JOURNAL OF EDUCATIONAL PSYCHOLOGY*, Vol. 9, May and June, 1918, 239-261; 333-348.) They are following directions, opposites, disarranged sentences, proverbs, arithmetic, geometric figures, analogies, similarities, narrative completion, and memory. Suitable for use in grades 5 to 12.

22. F. N. Freeman and H. O. Rugg. *The Chicago Group Intelligence Test*. School of Education, University of Chicago. A series of five tests (opposites, number completion, proverbs, analogies, and best reasons) based on the Army Alpha Tests, and designed for use in grades 5 to 12. It is said that this test is to be withdrawn.

23. W. W. Theisen and Cecile White Flemming. *A Group Classification Test*. Teachers College, Columbia University. Eight tests (following directions, synonym-antonym, arithmetic, common sense, completion, analogies, number completion, and information) for use in grades 5 to 12.

24. *Army Alpha Examination*, designed by a committee of the American Psychological Association for use with recruits during the war. The examination consists of eight tests (following directions, arithmetical problems, practical judgment, synonym-antonym, disarranged sentences, number series completion, analogies, and information), based largely on the Otis tests. These tests were widely distributed at the close of the war, but are now superseded by the National Tests.

25. Lewis M. Terman. *The Terman Group Test of Mental Ability*. World Book Company, Yonkers, N. Y. Ten tests (information, best answer, word meaning, logical selection, arithmetic, sentence meaning, analogies, mixed sentences, classification, and number series) are designed for use with grades 7 to 12.

26. L. L. Thurstone. *Psychological Examination for College Freshmen and High School Seniors*. Carnegie Institute of Technology, Pittsburgh, Pa. A series of 168 exercises, including infor-

mation, verbal relations, completion, proverbs, number completion, true or false statements, and logical reasoning, arranged in cycle formation.

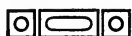
27. E. L. Thorndike. *Thorndike Intelligence Examination for High School Graduates*. Bureau of Publications, Teachers College, Columbia University. These are the well-known entrance examinations designed to supplement other entrance requirements at Columbia University.

28. David Camp Rogers. *Group Tests of Intelligence*. Published by the author at Smith College. There are 11 tests (logical conclusions, delayed recall of ideas, information, arithmetic problems, immediate recall of ideas, substitution, similar relations, completion, absurdities, following directions, and train of associations) designed for use with college freshmen.

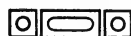
29. M. R. Trabue and F. P. Stockbridge. *The Mentimeter Tests*, in "Measure Your Mind: The Mentimeter and How to Use It." Doubleday, Page & Co., 1920. A series of thirty tests, derived largely from tests already in common use, and arranged roughly in ascending difficulty and complexity, so as to be applicable to all persons from infants to university graduates. In the preface the claim is made that "the Mentimeter is the first comprehensive system of tests, applicable to the whole range of educational and industrial requirements, to be offered for general use."

30. A. A. Roback. *Roback Mentality Tests for Superior Adults*. Prepared by the author for Simmons College. The Fenway, Boston, Mass., 1920. Twelve tests for college students. They are abstraction, problem, analogy, relations, insertion, reference, opposites, acumen, subsumption, directions judgment, and cryptogram. Total working time, 185 minutes.

DEPARTMENT FOR DISCUSSION OF RESEARCH PROBLEMS



Conducted by LAURA ZIRBES



This department has a two-fold function. It aims to serve research workers as well as educators, whose work brings them in close contact with children in the schools. It hopes to accomplish this service by suggesting research studies, which will meet well-defined school needs.

In order that this service may be real and effective, the co-operation of research workers and school people is desired. Correspondence with reference to the following questions will be considered in selecting topics for future discussions.

- a. Which of the studies proposed would help you to solve a practical problem?
- b. What topics might well be added to this list? Replies may be addressed to: Miss Laura Zirbes, 646 Park Ave., New York City.

In the last issue of this Journal one research problem was discussed at some length in this department, while five other possible studies were suggested. Because of the recency of our request for correspondence bearing on the suggested topics, it will be impossible to select problems for discussion from this source, as we hope later to do.

The following suggestions grew out of recent articles in educational publications, to which reference is made in each case.

Is the conscious co-operation of the child toward the attainment of known and concrete standards a factor which greatly affects improvement? Carefully controlled motivation of this sort over long and short periods. Control groups. Before and after measurement with comparative studies of improvement. Studies for several school subjects and grades. Results of incidental instruction compared with that above described.

S. M. Lloyd and C. T. Gray. University of Texas Bulletin No. 1853. Reading in a Texas City: Diagnosis and Remedy. P. 47.

What are the specific errors and difficulties of individual pupils in arithmetical learning and to what extent do they indicate that the necessary special bonds have been neglected? Careful accounting and analysis of all errors made by a class during a year. Lo-

cating the gaps in training and instruction and revising them to suit individual or class needs as revealed by analytical study of errors.

Edward L. Thorndike. The Constitution of Arithmetical Abilities. *Journal of Educational Psychology*, Vol. 12, No. 1, P. 14.

What are the limitations of Grade Standards and Age Norms in evaluating educational achievement? Constructive suggestions. Multiple Standards. New use of mental age. Achievement age.

That this problem is on the minds of a number of workers is evidenced by the number of articles in which it is suggested.

J. C. Bell. Editorial, *Journal of Educational Psychology*, Vol. XI, No. 4, P. 230.

B. R. Buckingham and Walter S. Monroe. A Testing Program for Elementary Schools. *Journal of Educational Research*, Vol. II, No. 2, P. 521.

Rudolf Pintner and Helen Marshall. A Combined Mental Educational Survey. *Journal of Educational Psychology*, Vol. XII, No. 1, P. 32.

To what extent would retardation and double promotion be diminished by admitting pupils to first grade when they are mentally six years of age regardless of chronological age? Would children of high and low I. Q.'s progress at similar rates? Are the current assumptions regarding mental growth valid?

Frank N. Freeman. Interpretation and Application of the Intelligence Quotient. *Journal of Educational Psychology*, Vol. XII, No. 1, P. 12.

NEW PUBLICATIONS IN EDUCATIONAL PSYCHOLOGY AND RELATED FIELDS OF EDUCATION*

1. *A book on educational tests for teachers.*—A book on educational tests and measurements, simple and direct in treatment, and practical in aim, for use of the classroom teacher has been needed. G. M. Wilson and Kremer J. Hoke¹ have done an admirable work to fill this need. The volume is not merely a collection of norms and descriptions of all available tests, with a scattering of statistical information and suggestions with regard to remedying deficiencies, nor is it replete with technical disputes. While it is admitted that “the tests which are going to survive and show value in the next few years cannot be determined at this time,” the authors have, in the case of most subjects, definitely recommended certain tests, for which directions for giving, scoring, and interpreting the results are given. The chief merits and deficiencies of other tests are briefly considered. For example, in treating spelling, the Ayres scale is recommended. A list of 20-40 words, selected from a list yielding about 70 per cent. correct spellings, should be dictated singly rather than in context. Definite, non-technical statements of computing, displaying and interpreting results follow, with some useful suggestions with regard to use of the data. Practically no space is devoted to discussions of methods of teaching, but three or four useful books on teaching of spelling are recommended. The merits and deficiencies of the Buckingham Scale, Buckingham’s Extension of Ayres Scale, the Iowa scale and the Rice, Starch, Boston, and Jones lists are briefly and effectively considered. The chapter is followed by a bibliography. In the case of handwriting, the authors favor the Ayres and Thorndike scales; in arithmetic, most attention is given to the Courtis, Series B, Stone Reasoning, Woody, Cleveland Survey, and the Boston Research Tests in fractions. Under reading, extended treatment is given to Thorndike’s Scale Alpha 2, Courtis Silent Reading Test No. 2, Monroe’s Standardized Silent Reading, Haggerty’s Visual Vocabulary, and Gray’s Oral Reading Test, while other tests are briefly considered. The Nassau County Supplement

¹Wilson, G. M. and Kremer J. Hoke. *How to Measure*. New York: Macmillan, 1920 Pp. VII + 285.

and Willing Scales are described in the chapter on English composition, and directions for making up a local scale are given. A short chapter is devoted to Thorndike's Drawing Scale, another, to scales in history, geography, language, and music, and another, to high school subjects.

The inclusion of a chapter on the measurement of general intelligence is a most useful innovation. The elements of mental status and its bearing upon school attainment are briefly presented, and the teacher is instructed how to use the Trabue Language Scales, the Otis and the Haggerty tests as measures of intelligence. We are surprised that the National Tests were not included, and the treatment of the Stanford-Binet test is clumsy. It is unfortunate that this chapter is the least satisfactory in the book.

The chapter on statistical terms and methods is short. A merit of the volume is the avoidance of statistical terms and procedures, which, more than anything else, keep teachers from making friends with educational tests and scales. The final chapter deals with methods of grading pupils; the unreliability of conventional methods, and a dozen special advantages of standardized tests are given.

"We are now quite surely approaching a third stage of development, and that is the state in which the tests shall be thoroughly weighed and judged as to the fundamental considerations of curricula making involved, whether they are or are not testing desirable school products, and whether their use will or will not lead to better methods of teaching and better selection of subject-matter. In this stage, the standard tests will be used more and more for diagnosis of the weaknesses of individual pupils, more and more in testing the efficiency of methods of teaching. It is in this third stage that the rank and file of the teaching profession are necessarily involved."

This book will be of service in this movement.

A. I. G.

2. *A semi-popular book on intelligence.* It is a rare treat when a recognized authority tries, and succeeds, in describing the results of scientific work in popular form. Dr. Goddard² has succeeded in his small volume on human efficiency as determined by levels of intelligence. It will afford useful and interesting reading to a wide public who wish the outstanding facts and implications of mental testing, in a form devoid of technical devices and vocabulary. The

²Goddard, Henry Herbert, *Human Efficiency and Levels of Intelligence*. Princeton: University Press, 1920. Pp. VII + 123.

first chapter considers the nature of intelligence, the methods of testing, the distribution of individuals, and uses that have been made of results. The results of intelligence testing in the army is summarily surveyed. Chapter two shows the influence of intelligence levels in industrial, educational, and social adjustment. A third chapter deals with delinquency and crime. As regards causes, the author cites an answer to the question, "Why does a child go wrong?" "Either because he does not know any better, or because he cannot help it." "The new thought in this connection is that we have only to extend these two principles in order to account for practically all of juvenile delinquencies and a large part of adult criminality." In addition to the influence of low mentality, mental diseases and "pure wickedness" as causes of crime, the treatment and investigation of treatment of criminals, bases upon mental status is suggested, with an account of plans of the Iowa Bureau of Juvenile Research in this connection. The final chapter, the relation of individual differences in intelligence to a democratic form of government, is discussed. "The intelligent group must do the planning and organizing for the mass, and our attitude toward the lower grades of intelligence must be philanthropic." A small percentage should be given institutional care, working at tasks within their limitations, and throughout the nation, educational and vocational adjustment should be made in terms of intelligence levels. The effects of such placement will be an increase in efficiency and happiness of the people, a reduction of delinquency, crime, prostitution, reproduction of the unfit, and social maladjustment. "Each mentality requires its own kind of life for its success and happiness."

If one does not keep in mind Dr. Goddard's preface, to the effect that "it seems worthwhile to solve our problems in terms of intelligence, as though it were the only variable," because the role of other factors is not well understood, one would feel that the influence of intelligence is exaggerated. The influence of emotional and temperamental traits, home and school advantages, etc., are considered largely as subsidiary. "While intelligence or mental level is not the sole factor in human efficiency, it is, nevertheless, the determining factor, and our social inefficiency is due primarily to the large percentage of low intelligence, and, secondly, to a lack of appreciation of relatively low intelligence by those of higher intelligence." P. 57.

The discussion of intelligence levels and school education is brief. Among the superior children are two groups: first, those who are truly gifted children, and, second, those whose brilliancy is evidently due to a very high-strung nervous system, cases of the "well known, but little understood relationships between genius and insanity." The reader will be disappointed to find no extensive discussion of these types, most emphasis, necessarily, no doubt, being given to the sub-normal groups. The choice of education and vocations for the latter is outlined. The educational guidance by early tests and selection by tests at high school and college entrance is endorsed. Finally, "Why should we not ascertain the grade of intelligence necessary in every essential occupation, and then entrust to that work only those people who have the necessary intelligence?" P. 118.

Technical matters, such as the validity of the Intelligence Quotient, are omitted. Certain charts and statements give the impression that the writer considers all individuals to have a constant growth, and "that *at any age* development may cease and the individual remain at that level the rest of his life." Whether this is literally meant could scarcely be determined by a reader unacquainted with other concepts on the field. That "the mental level is determined with marvelous accuracy by trained psychologists," and that "the results of the army testing is of profound significance," illustrates the optimistic spirit in which the book is written. It is, however, a very clear and most interesting account of the results and possibilities of intelligence testing.

A. I. G.

3. *A new textbook on fundamental principles of learning and study.*—Professor A. S. Edwards, University of Georgia, has written a book³ directed primarily to college students of education, which is adapted as well to school principals and supervisors, and parts of it may be studied with profit by high school students. Accepting the word "habit" to express the acquisition of memories, skills, interests, and "all more or less permanent tendencies of mind and body," the author proceeds to discuss most of the topics within the field of learning. In general, the thesis is, that while all learning is habit formation, it should be of permanent value, and generalized improvement, as well as specific improvement, should be sought. The acquisition of ideals, habits of efficient thinking, working, memorizing, studying, of emotional control et al., are discussed in

³Edwards, A. S. *The Fundamental Principles of Learning and Study*. Baltimore: Warwick and York, 1920. Pp. 240.

turn. While the general principles are not most economically handled, the repetition in the application to the several types of mental work may be of advantage to the student. The book abounds in definite practical directions; two chapters in particular consist almost wholly of such admonitions as "Try to find an application for everything you learn," "Review frequently," "Have an intention to remember," "Learn at your own best rate," etc.

The book is neither a compendium of experimental studies on learning, nor a novel systematic treatise on principles. Quotations in abundance from James, Meumann, Thorndike, Judd, Sully, Angell, Carpenter, and others are included, but a great deal of experimental work of merit has not been drawn upon. It will be difficult for the student to distinguish principles based on experimental findings from the opinions of the author. However, it is a useful collection of suggestions with regard to methods of study, and in many respects has merits not possessed by popular books on "How to Study." Suggestions with regard to appropriate diet, humidity, temperature, change of work, sleep, naps, light, and the like are included.

A. I. G.

4. *A report on the diagnosis and improvement of pupils' reading ability in a city school system.*⁴—Although the author's preface indicates that this bulletin was prepared especially for the use of elementary teachers of Texas, it will, no doubt, find a larger field of use. The entire literature of diagnosis and remedy as listed in chapter III and in the Bibliography is so meagre that it is well supplemented by a careful study of this sort.

From the available Reading Tests described in the first chapter two were chosen and used as described in chapter two.

The program of diagnosis and remedial work is described in detail in chapters four and five. The numerous charts and diagrams should make it quite possible for any teacher to make regular use of the procedures devised. That such use would prove profitable is indicated by the re-measurement and interpretation of final scores in tests given after four months of remedial work.

The final chapter on "Summary and Suggestions" gives illustrative records and practical suggestions of the sort which makes it possible

⁴Lloyd, S. M. and Gray, C. T., *Reading in a Texas City: Diagnosis and Remedy* University of Texas Bulletin No. 1853.

for any average teacher to use the carefully planned technique. This is, no doubt, the greatest contribution of the study.

L. Z.

5. *A college textbook on the social significance of human traits.*—"To give a bird's-eye view of the processes of human nature from man's simple unborn impulses and needs, to the most complete fulfillment of these in the deliberate activities of religion, art, science, and morals," is the rather large task accepted by a philosophic writer.⁵ What we find is a rather loquacious discussion of innumerable aspects of human behavior, with a great deal of repetition. In Part I, 275 pages are devoted to "social psychology," in which man's instinctive equipment, prolonged infancy, the origin of language, the mechanism of habits, learning by trial and error, "deliberate learning," drill, fatigue, health, social inertia, conscious transference of habits, reflection, opinion, belief, character, will, the development of the self, individual differences, language and communication, racial and cultural continuity, and many other things are discussed. Part II devotes nearly 200 pages to the "Career of Reason" in the case of religious, artistic and aesthetic experiences, science and scientific methods, morals and moral valuations. The influence of the psychological writings of Thorndike, Woodworth, and McDougall, and the philosophical works of James, Dewey, and Santayama appears throughout. For the most part, the book attempts to explain and apply the principles advocated by these writers. There is but little new material that will be of interest to psychologists, and the book is not adapted specially to the needs of education. It will be found most useful as a text for an elementary course, combining psychology and philosophy, having been written primarily for use in a course entitled "Introduction to Contemporary Civilization," required of freshmen in Columbia College. It may be enjoyed by the mature general reader who lacks the time to read the several books from which the present volume is largely drawn.

A. I. G.

6. *A study of women delinquents in New York.*—Probably the most intensive study of delinquency among women in described in

⁵Edman, Irwin. *Human Traits and their Social Significance*. Boston: Houghton Mifflin. 1920. Pp. XIX + 467.

this large book,⁶ written jointly by Mabel R. Fernald, Mary H. S. Hayes, Almena Dawley, and Beardsley Ruml. Some 550 cases were studied by use of the Yerkes-Bridges, Stanford-Binet, the Woolley general intelligence tests, educational performance, and other tests by a staff of psychologists. Social investigators secured family histories, school records, hospital, institutional, criminal, and vocational histories, and a detailed account of circumstances preceding and following offenses. Information concerning habits of addiction to drugs, alcohol, and tobacco was secured. The most important findings relate to educational attainment, mental levels, early home conditions and occupational history and efficiency. As regards educational attainments, the mean grade reached is 4.5, with about 15 per cent finishing the eighth grade and 1.6 per cent finishing high school. Eleven per cent had received no regular schooling. Educational tests showed a group median about equal to that of the fourth grade. The intelligence rating by the Stanford-Binet gave a mean of 11.8 years, which may be compared to 13.4 years, the mean of an army group. Delinquent women apparently are, as a group, somewhat inferior to the general public, in both intelligence and school attainments, but the overlapping is so great, and the increase of the delinquent tendencies so slight with a decrease in intelligence, that other important causes must be sought. The results of the social and economic histories and surveys led the authors to the following conclusion as regards causation: "A somewhat inferior intellect," coupled with poor economic background, with few advantages and opportunities, including such conditions as poor homes, bad company, limited schooling, early age of starting to work, and no useful industrial training at length becomes involved in economic straits from which prostitution, theft, etc., offers a most convenient escape. More than many early writers, especially psychologists, the authors are inclined to emphasize the role of environmental influences, rather than constitutional limitations, as the cause of delinquency.

A. I. G.

7. *A book on physical growth by a French writer.*—Mr. S. L. Eby states that his purpose in translating Dr. Godin's book from the French is, first, to introduce the writings of a Frenchman who has

⁶Fernald, Mabel R. Hayes, Mary H. S. Dawley, Almena, and Ruml, Beardsley. *A Study of Women Delinquents in New York State*. New York: The Century Co., 1920. Pp. XVIII + 541.

long been a student of scientific education, and, second, to direct greater attention to the contributions to the theory and practice of education in France.⁷ Dr. Godin's point of view is admirable:

"The only physical measurements worth while are those which admit of comparisons with previous states of development of the same individual. Such comparisons can be valid only when repeated measurements are taken at regular intervals. These repeated measurements are necessary in order to enable the teacher and educator to know the child intimately and profoundly; it makes possible a degree of individualization of education unknown in the past."

In order not to lose the variations in "rhythm" and alternation in growth, measures are taken semi-annually. Emphasis is placed upon the fact that gross measures mean little unless broken up by finer analysis. Growth is not uniform, but proceeds with variations in the case of the limbs, trunk, and head. The book includes many charts and diagrams illustrating segmental development from birth to puberty.

Part I is devoted to a discussion of the nature of growth, and Part II applies the results to education. No actual data are given, a fact which makes the volume seriously unsatisfactory. As regards applications to education, many of them are most curious, for example, the following:

"The trunk is cubed by the product of the multiplication of its dimensions. * * * The product of this double multiplication is called V (Viscera). V varies enormously from birth to adult age, and it is expressed by a different figure at each of the stages of growth. * * * It is the same with the product of C of the double multiplication of the diameters of the cranium, of which the content is the encephelon, the brain, consequently. The relation of C to V gives a quotient which instructs us on the relative proportions of the viscera, of the vegetative life, and of those of the psychic life. The quotient informs the educator of the free field which the individual vegetative resources for cerebral culture leave to him."

A. I. G.

⁷Godin, Paul. *Growth during School Age; its Applications to Education*. Translated by Samuel L. Eby. Boston: Richard G. Badger. 1920. Pp. 268.

8. *A condensed guide for the Stanford Revision of the Binet Tests and Abbreviated Filing Record Card.*—Professor Terman⁸ has published a guide for use of his revisions of the Binet tests which is convenient to carry and arranged for fluency of reading in that the material to be spoken to the subject is printed in black-face type, while the notes of interest to the examiner are printed in light type. No significant changes in procedure appear. All the essentials for giving and recording the tests are now printed on a single card,⁹ typewriter size, which is economical of space and more convenient for filing.

A. I. G.

9. *A monograph which evaluates the present status of home economics in our schools.*—The first step in the reconstruction of a curriculum is the thorough evaluation of the existing situation. In the past five years much of this evaluative work has been done in elementary and junior high school subjects. Careful reports have been made on reading, handwriting, spelling, arithmetic, and ninth-grade mathematics. Many unpublished analyses are to be found on the shelves of our educational libraries. To the present time, no such evaluation has been available for home economics.

A series of careful studies in this field has just been published in one of the "University of Chicago Supplementary Educational Monographs."¹⁰

These studies were made under the direction of Dr. H. O. Rugg by members of the departments of home economics in the University of Chicago and in the Iowa Agricultural College.

The surveys include a quantitative study of existing courses in elementary and secondary schools; a minute analysis of the textbooks, by which so much of the teaching is controlled, and a careful canvass of the literature of the subject in search of definite aims and objectives; in addition, a preliminary organization of tests for measuring skill in machine sewing, tests for content in textiles and clothing, and a suggestive program for the reconstruction of the curriculum in home economics.

The authors report that a thorough examination of 67 of the city public school courses shows that "home economics" means in nearly

⁸Terman, Lewis M. *Condensed Guide for the Standard Revision of the Binet-Simon Intelligence Tests*. Boston: Houghton Mifflin Co. 1920. Pp. 32.

⁹Terman's *Abbreviated Filing Record Cards*, published by Houghton, Mifflin Co., Boston.

¹⁰Rugg, H. O., in collaboration with the departments of home economics in the University of Chicago and in the Iowa Agricultural College, "*Home Economics in American Schools*." Chicago: University of Chicago, 1920. Pp. X + 133.

all cases merely cooking and sewing. The courses are organized formally, with practically no attention to subject-matter sequence or to learning difficulty. Conclusions concerning the organization of home economics curricula were derived by evaluating the course against the four following psychological criteria:

1. "To what extent does the home-economics curriculum provide for worthwhile home-making 'skills'?"
2. "Is the time of children in home economics classes devoted to acquiring information of social value?"
3. "Is sufficient opportunity provided for young people to develop powers of critical judgment—that is, analytical thinking?"
4. "Are we so planning our curriculum and organizing our class exercises in home economics that there is promise that the instruction will eventuate in real ability to appreciate and enjoy?"

On the basis of such definite measures, and on definite principles of selection and arrangement of subject-matter, the writers report that, "information giving and the development of technique dominates the course," and that the primary importance of training children in judgment and appreciation is lost sight of. An investigation shows by a tabular analysis of the 19 most frequently used text books that the books and the courses are so encyclopedic and so unpsychologically arranged that it may fairly be said that children's learning is inhibited rather than promoted.

Purposes are scrutinized in five sources in the literature, in the same objective way, and with similar conclusions: that purposes are inadequate as revealed by the emphasis on the mastery of information and technique. The report follows this sweeping criticism by a constructive statement of new purposes for teaching: that is training power of thinking and enjoyment.

The writers next show that of two the effective methods by which a sound scheme of instruction in home economics can be organized, the first is the statement of outcomes; that the second is the design and use of standardized measures of instruction. They report new tests for results in textiles, dress designs, sewing, and house-planning, and new scales for measuring skill in machine sewing. Finally, the report presents a program for improvement through the scientific study of home economics education.

10. Professor Hanus has just published in book form¹¹ a series of essays, largely in the field of school administration, which have appeared in various educational periodicals during the past nine years. The volume is intended to "help the superintendent of schools, and other persons who are charged with the responsibility of providing good schools and school systems for the public, to formulate and justify their opinions and procedure."

The material that is primarily administrative is found in the first three essays. These deal critically with the problem of educational aims in a school system; with a definite statement of fundamental principles, which every superintendent who seeks to define his administrative policy needs to assimilate; and with the aims, scope, and method of town and city school reports. The fourth essay directs at the superintendent questions concerning the efficiency of his school system, which it would be well if each school executive applied perennially to his own schools.

Professor Hanus's activity in the scientific field in education is reported in two essays, one dealing with his use of the Courtis Arithmetic Tests with business employees, and the other his pioneer attempt to devise a test for ability in high school Latin. Another essay treats with the field of state school administration.

The volume closes with three essays which critically compare educational aims and practices in Germany and the United States.

III. ADDITIONAL PUBLICATIONS RECEIVED.

A. MENTAL AND EDUCATIONAL TESTS.

HAGGERTY, M. E. *Reading Examination*. Sigma 1 and Sigma 3. Yonkers: World Book Co., 1921.

HOLLEY, CHARLES E. *Mental Tests for School Use*. Urbana: University of Illinois, 1920. Pp. 6 + 91.

HUDELSON, EARL. *Hudelson English Composition Scale*. Yonkers: World Book Co., 1921. Pp. VII + 46.

SEASHORE, CARL E. *A Survey of Musical Talent in the Public Schools*. Iowa City: University of Iowa, 1920. Pp. 36.

VAN WAGENEN, MARVIN J. *Historical Information and Judgment in Pupils of Elementary Schools*. New York: Teachers College, Columbia University, 1919. Pp. 74.

WILLIAMS, J. HAROLD. *A Survey of Pupils in the Schools of Bakersfield, Cal.* California: Whittier State School, 1920. Pp. 43.

¹¹Hanus, Paul H. *School Administration and School Reports*. Boston: Houghton, Mifflin Company, 1920. Pp. XI + 200.

B. PUBLICATIONS IN THE GENERAL EDUCATIONAL FIELD.

- COLLEGE TEACHERS OF EDUCATION. *Studies in Education*. Iowa: Tribune Publishing Co., 1920. Pp. 32.
- JOHANSEN, F. O. *Projects in Action English*. Boston: Badger, 1920. Pp. 207.
- NORTH CAROLINA: STATE EDUCATIONAL COMMISSION. *Public Education in North Carolina*. New York: General Education Board, 1920. Pp. X + 137.
- REED, ANNA Y. *Junior Wage Earners*. New York: MacMillan Company, 1920. Pp. XII + 171.
- STOCKTON, J. L. *Project Work in Education*. Boston: Houghton Mifflin Company, 1920. Pp. XIV + 167.
- TRUEMAN, G. J. *School Funds in the Province of Quebec*. New York: Teachers College, Columbia University, 1920. Pp. 154.
- WOODY, THOMAS. *Early Quaker Education in Pennsylvania*. New York: Teachers College, Columbia University, 1920. Pp. 287.

C. NEW SCHOOL TEXTBOOKS.

- FOX, D. R. *Atlas of American History*. New York: Harper & Bros., 1920. Pp. 181.
- GRISCOM, E. *Americanization*. New York: MacMillan Company, 1920. Pp. 255.
- TRAUSEAU, E. N. *Science of Plant Life*. Yonkers: World Book Co., 1921. Pp. IX + 336.
- SAIT, E. M. *Government and Politics of France*. Yonkers: World Book Co., 1920. Pp. XV + 478.
- WASHBURN, C. W. *Common Science*. Yonkers: World Book Co., 1920. Pp. XV + 390.

D. PUBLICATIONS OF UNITED STATES BUREAU OF EDUCATION.

- BONNER, H. R. *Statistics of Public High Schools 1917-1918*. Bulletin 19, 1920. Pp. 192.
- BONNER, H. R. *Statistical Survey of Education, 1917-1918*. Bulletin 31, 1920. Pp. 48.
- BONNER, H. R. *Private Commercial and Business Schools, 1917-1918*. Bulletin 47, 1919. Pp. 123.
- COMMISSIONER OF EDUCATION. *Annual Report*. 1920. Pp. 134.
- Home Economics Courses of Study for Junior High Schools. Home Economics Circular 9, 1920. Pp. 14.

E. MISCELLANEOUS PUBLICATIONS.

- BAMESBERGER, V. C. *Standard Requirements for Memorizing Literary Material*. Urbana: University of Illinois, 1920. Pp. 93.
- GALLOWAY, T. W. *The Sex Factor in Human Life*. New York: The American Social Hygiene Association. Pp. 56.
- Preliminary Synthesis and Integration of the Returns of the Sex Education Conference*. New York: The American Social Hygiene Association. Publication 321. Pp. 95.
- STINCHFIELD, S. M. *Preliminary Study in Corrective Speech*. Iowa City: University of Iowa, 1920. Pp. 36.

THE JOURNAL OF EDUCATIONAL PSYCHOLOGY

Vol. XII

MARCH, 1921

No. 3

INTELLIGENCE AND ITS MEASUREMENT: A SYMPOSIUM

Editorial introduction: Probably the most striking advance of our generation in the practical application of psychological technic to educational and social affairs is the movement for the definition and measurement of intelligence. Especially impressive is the widespread effort now devoted to the construction and use of individual and group tests. It appears that this is an opportunity for this journal to serve as an effective clearing house for mature opinion on a most important problem. Accordingly, we have asked 17 leading investigators to contribute to a symposium on the following topics:

- “(1) What I conceive ‘intelligence’ to be, and by what means it can best be measured by group tests. (For example, should the material call into play analytical and higher thought processes? Or, should it deal equally or more considerably with simple, associative, and perceptual processes, etc.?)
- “(2) What are the most crucial ‘next steps’ in research?”

Those invited to take part were: Doctors Bell, Buckingham, Colvin, Dearborn, Freeman, Haggerty, Henmon, Peterson, Pintner, Pressey, Ruml, Terman, Thorndike, Thurstone, Whipple, Woodrow, Yerkes.

Of these, all but three have signified their desire to contribute material. We print in this issue statements by Drs. Thorndike, Terman, Freeman, Colvin, Pintner, Ruml, and Pressey. Contributions from the other workers will be published in the April issue.

The editors hope that a vigorous discussion will grow out of these initial statements and will present further phases of the matter in subsequent issues of the Journal.

I. By E. L. THORNDIKE,
Teachers College, Columbia University.

1. *The nature and measurement of intelligence.* If we inventory the behavior of men and try to divide it up according as intellect, character, skill, taste or temperament is primarily involved, we shall agree fairly well in, say, ninety per cent of the cases. If, however, we try to make the division absolute we may agree very seldom. It is probably unwise to spend much time in attempts to separate off sharply certain qualities of man, as his intelligence, from such emotional and vocational qualities as his interest in mental activity, carefulness, determination to respond effectively, persistence in his efforts to do so; or from his amount of knowledge; or from his moral or esthetic tastes. Even so apparently remote a trait as muscular strength may in some cases cooperate almost indistinguishably in the production of what we would all call intellectual products. Thus a great scholar's achievement may be in part due to eye muscles which help make reading a pastime.

Taking these cases of behavior which are generally accepted as matters of intellect and trying to place each as primarily a matter of response to situations directly sensed, or as primarily a matter of planning, we shall again agree fairly well. So also if we rate them as primarily responses to concrete particulars or primarily responses to abstract qualities and relations. It would however be difficult and probably unwise to try to separate off *sharply* the responses concerned with directly sensed situations from responses concerned with planning; or those concerned with concrete features of things and men from those concerned with ideas and symbols. Hoeing corn and shooting a rabbit are easily distinguished from studying botany and ballistics, but behavior shows all sorts of intermediate forms.

Realizing that definitions and distinctions are pragmatic, we may then define intellect in general as *the power of good responses from the point of view of truth or fact*, and may separate it according as the situation is taken in gross or abstractly and also according as it is experienced directly or thought of. The power of good responses to abstract qualities and relations rather than gross total facts and to ideas rather than direct experiences may be called the more intellectual variety of intellect.

type
intelligence
d
ly

1
D

Either variety is, as human beings now are, specialized further according to the data operated on in the response and according to the form of the operation. The goodness of the response in any individual varies according to the particular task. The child who is the best of a thousand of his age at the undoubtedly intellectual task of mental multiplication of two-place numbers will not be the best at the equally indubitably intellectual task of thinking out verbal puzzles. As we change the task from space relations to animals, to plants, to machines, to moral issues, to numbers, the correlation never holds up to 1.00. As we change from accuracy in perception, to accuracy in memory, to accuracy in inference, all with the same data, the correlation again fails to hold at 1.00.

A part of this break to below perfect correlation is presumably due to circumstances of life and training which have given unlike amounts of emphasis to different data and to different forms of operation in the case of each of the thousand. But if a thousand were taken who had had identical training, it seems certain that the specialization of intelligence would still be found, the correlations still failing to be unanimously 1.00.

In measuring a person's general status in intelligence and in inferring therefrom what his rank in native intellectual capacity in general is, what we do is to test him with a fair sampling of data and operations. If his opportunities of training in respect to these have been inferior or superior to the group with whom he is to be compared we make the necessary allowance. This sampling should be wide enough and its various components should be easily enough weighted, so that the resulting judgment should be about his *general* status and *general* capacity—if we are to claim that it is general.

Some of us have, I fear, claimed a generality for our measures of status and a surety of inference from them to original inborn capacity which it would be very hard to justify. The estimates which the psychologist makes with his tests are much better than those which parents or teachers or ordinary medical practitioners make of the same facts, so that we are justly proud of them, but we should be the first to recognize their limitations. The value of a test score is its value in prophesying how well a person will do in other intellectual tasks. Our claims may wisely be limited to the actual demonstrated power of prophecy. For example, consider a score attained by a 12-year-old boy in say a combination of Stanford Binet, National A and

B and Haggerty Delta 2 (two trials of each). If the boy has had ordinary American opportunities, this score will prophesy rather accurately how well he will respond to intellectual demands in the cases of "book-learning" at the time and for some years thereafter, and very possibly for all his life. It will prophesy less accurately how well he will respond in thinking about a machine that he tends, crops that he grows, merchandise that he buys and sells and other concrete realities that he encounters in the laboratory, field, shop and office. It may prophesy still less accurately how well he will succeed in thinking about people and their passions and in responding to these. We know that, taking people as we find them, the ability measured by verbal tests is not the same as the ability measured by non-verbal tests; and there is reason to expect other similar specializations.

The intercorrelations of these various "intelligences" are, of course, high enough to make a measure of any one of them a better index of any other than the average parental hope or teacher's opinion is likely to be. If our 12-year-old boy is at the five-percentile station in our tests it will rarely happen that he will rank above average intelligence over any large area of mental activity; save by the drive of a great interest and the expense of much time upon the activity in question. On the other hand to assume that we have measured some general power which resides in him and determines his ability in every variety of intellectual task in its entirety is to fly directly in the face of all that is known about the organization of intellect.

The relative weight to be attached to analytical and selective and to perceptual and simple associative processes can only be decided by the correlations and partial correlations of these with the criterion which the test score is to prophesy, or some experimental facts giving equivalent information, considered with due reference to the time-cost of making the prophecy. In general, tasks which require efficiency in analyzing a situation into elements, selecting and weighting elements to fit a problem and organization or thinking many things together seem to give much better results per dollar or hour of cost. They perhaps include and sum up the action of many simpler associative processes. But a straightforward information test is also a valuable element. If time-cost is disregarded, a sampling of every kind of intellectual operation, and with every kind

of *datum*, will, if properly weighted, improve the prophecy of *general* intelligence. Unless it is properly weighted, however, it may injure it.

2. *Next steps in research.* Having used up my allotment of space in the previous discussion, I can only note that more research into every feature of intelligence and its measurement is needed, especially researches on the intercorrelations "total" and "partial" of the various forms of intellectual work of the world; on the intercorrelations "total" and "partial" of the abilities used and usable in our tests; on the permanence of our IQ's and EQ's; and on the "total" and "partial" correlations of each test at each age with itself at later ages and with each feature of the intellectual work of the world. The form of distribution of intelligence in the general adult population is also a matter of great practical moment. In relation to educational and industrial uses the problem of the effect of 'coaching,' special and general, is fundamental.

II. By L. M. TERMAN,
Leland Stanford University.

1. *The nature and measurement of intelligence.* Meumann has pointed out that the fault of Stern's teleological definition of intelligence as "general adaptability to the new problems and conditions of life," lies in the fact that it furnishes no clue for judging the value of different kinds of adaptation. Meumann would reverse Stern's procedure by first finding out what is demanded of intelligence and then analyzing the mental functions which meet that demand. In my opinion this is the only method of approach which will bring us any nearer to a psychological solution of the intelligence problem.

If we accept this view it is evident that the important intellectual differences among men will not be found on the sensory, perceptual, or purely reproductive level. It is well known that a moron may be able to see, hear, taste or smell, react to a signal, balance a bicycle, steer an automobile, or cancel A's about as well as an intellectual genius. The latter would be somewhat his superior in memory for non-sense syllables, would excel him still more in logical memory, and would outclass him hopelessly in the ability to distil meanings

from the raw products of sensation and memory. The essential difference, therefore, is in the capacity to form concepts to relate in diverse ways, and to grasp their significance. *An individual is intelligent in proportion as he is able to carry on abstract thinking.*

One may, of course, question our grounds for designating any kind of mental activity as "higher" or "lower" than another. Why, it may be asked, should certain types of mental processes be singled out for special worship? In fact, it is frequently intimated that the individual who flounders in abstractions but is able to handle tools skillfully, or play a good game of baseball, is not to be considered necessarily as *less* intelligent than the individual who can solve mathematical equations, acquire a huge vocabulary, or write poetry. The implication is that the two individuals differ merely in having different *kinds* of intelligence, neither of which is higher or better than the other.

It is difficult to argue with anyone whose sense of psychological values is disturbed to this extent. Such an individualistic view is not disposed of by calling attention to the obvious fact that civilization, with its science, art, government, religion, philosophy, and systems of credit, is unthinkable except as a product of concept elaboration and symbolic thinking; our opponent can retort that it is only our intellectual snobbery which leads us to regard the state of so-called civilization as "higher" than that of primitive man!

It can not be disputed, however, that in the long run it is the races which excel in abstract thinking that eat while others starve, survive epidemics, master new continents, conquer time and space, and substitute religion for magic, science for taboos and justice for revenge. The races which excel in conceptual thinking could, if they wished, quickly exterminate or enslave all the races notably their inferiors in this respect. Any given society is ruled, led, or at least molded by the five or ten per cent of its members whose behavior is governed by ideas. The typical pick-and-shovel man does his thinking chiefly on the sensori-motor and perceptual levels. Add a little more ability to think on the representative level and he may be able to repair your automobile, build you a house according to an architect's specifications, or nurse you in illness. Add a large measure of ability to associate abstract ideas into complex systems and he can design a new type of engine, draft the plans for a skyscraper, or discover a curative serum.

What I have said may seem like an over-elaboration of the obvious. Such a discussion ought, indeed, to be unnecessary. Unfortunately it is not. Many criticisms of the current methods of testing intelligence rest plainly on a psychology which fails to distinguish the levels of intellectual functioning or to assign to conceptual thinking the place that belongs to it in the hierarchy of intelligences. If an intelligence test can be shown to depend upon the language factor (i. e., upon the ability to think in terms of symbols), this is sufficient in the eyes of some psychologists to condemn it as non-valid. The subject who can not acquire a normal vocabulary, see the point of a fable, or be taught to read a paragraph with understanding is considered to have demonstrated his intelligence if he can trace a simple maze or assemble the fragments of a formboard. In fact the "idea thinker" is sometimes spoken of disparagingly or even a little contemptuously, particularly in the case of a child whose superior ability in this respect places him in conspicuous contrast with other children of the same age.

But if intelligence is the ability to think in terms of abstract ideas, we should expect the most successful intelligence tests to be just those which involve the use of language or other symbols. We should also be justified in demanding that an intelligence test should correlate well with what we may call "school educability." As a matter of fact it is precisely tests of this type which are surviving in the struggle for existence: tests involving arithmetical reasoning, language completion, naming opposites, matching proverbs, completing analogies, understanding difficult passages, etc. The list could of course be greatly extended, but all tests which are notably successful in measuring intellectual differences among adults have something in common with those named, even when they work with concrete materials instead of with words.¹

With this conception of intelligence we can understand why it is so difficult to devise tests of the non-verbal or "performance" type which will bring out intellectual differences much above the level of the average child of ten or a dozen years. That it *is* difficult must be admitted by anyone who has attempted it and checked up his results. Non-verbal tests of the type used in the army Beta and performance examinations have of course proved useful on the intermediate levels of mental ability. However, even for children of

¹It is not implied, of course, that an ideal intelligence scale would necessarily be composed entirely of verbal tests.

eight or ten years, no purely non-verbal battery of tests yet devised is as successful as any one of several existing batteries which make some demands upon the language factor. Tests for very young children must of course be still simpler and work largely on the perceptual or at best on the lower representative and reproductive level. With very young infants, intelligence tests, as such, are of course not applicable. The best we can do here is to determine the presence of a nervous organization which will later make intelligence possible (e. g., tests of ability to acquire conditioned reflexes).

That is, there is no type of test which will measure intelligence equally well at all levels. It is largely due to Binet's recognition of this fact that his intelligence scale has proved so successful. Instead of trying to measure supposed linear faculties, or "functions," Binet undertook to determine levels of mental activity. (The reader may have observed that the "faculty" devils so ruthlessly cast out of the psychological temple by Herbart and James are wont still to parade among us under the alias of "mental functions.") He recognized that with increasing maturity intelligence becomes something different. Ruml's recent criticism of current test methods on the ground that these methods are usually based on the assumption that general intelligence can be expressed as a linear or one-dimensional function, was antedated some fifteen years by the following statement of Binet: "This scale (Binet's 1905 series of tests), properly speaking, does not permit the *measurement* (*italics added*) of intelligence, because intellectual qualities can not be measured as linear surfaces are measured, but are, on the contrary, a classification, a hierarchy among diverse intelligences."¹ While all will admit that in occasional overdrawn statements Binet sometimes carries his theory of levels too far, I do not know of any other psychologist who has brought such penetrating insight and such power of psychological analysis to bear on the intelligence problem.

It is clear why no intelligence scale made up of a battery of serial tests, each intended to give a linear measure of some supposed mental "function," has ever worked successfully over a very wide range. Even in the relatively narrow range in which it operates, such a test does not, at all points, bring the same kind of mental activities into

¹In the above reference to Dr. Ruml's criticism of mental test methods the writer does not wish it to be understood that he agrees with the general spirit of the article in question or with all of its criticisms, some of which he believes to be ill-founded and unjust.

play. Success in the easier part of the test may depend chiefly on the subject's ability to remember what he is told to do; see likeness and differences on the representative level, or even to a considerable extent on eye-hand coordination in the use of a pencil. In its more difficult parts the same test may have to do with the most subtle types of relationship among highly abstract ideas. How naïve to suppose that each test of such a battery measures a particular function! How absurd, also, to regard intelligence of every kind as of equal rank with intelligence of every other kind!

In closing I should like to point out two pitfalls which ought to be avoided in our thinking about the nature of intelligence. They are opposing dangers, and if we are overcautious in avoiding the rock of Scylla we are so much the more likely to be drawn into the whirlpool of Charybdis. (1) We must guard against defining intelligence solely in terms of ability to pass the tests of a given intelligence scale. It should go without saying that no existing scale is capable of adequately measuring the ability to deal with all possible kinds of material on all intelligence levels. Accordingly, the validity of a new test should not be judged entirely by its correlation with existing tests, however good these may be. There must be continued search for useful outside criteria. (2) On the other hand, in our anxiety to escape the evils of a closed system we must guard against indiscriminate and ill-considered use of outside criteria. To condemn an intelligence test because it yields low correlations with success as a mill hand or street car motorman is an example of this error. Another mistake that results from over-evaluation of a single criterion is seen in the effort to embody in a given intelligence scale *every* kind of test which will add to its correlation with the criterion in question. This mistake is especially likely to occur if one is interested in the predictive uses of the test. The effect of this error may be to pervert grossly the test as a measure of intelligence. If we wished to devise a test which would give the most accurate possible prediction of the class marks a given group of college students would receive, we ought to include in it measures of personal beauty, voice quality, bashfulness, willingness to cultivate the good graces of the instructor, etc. All of these qualities undeniably influence a student's marks, but they do not belong in a battery of tests designed to measure intelligence. It should not be the aim of an intelligence test to give us the *best possible* prediction of events of this kind.

Let us use outside criteria, by all means, but let us use them with psychological discrimination.

2. *Next steps in research.* I would list the following as a few of the many important "next steps" in intelligence testing. No significance is to be attached to the order of mention.

1. Systematic exploration for valid tests of particular aptitudes, including abilities especially concerned with success in typical kinds of vocational employment and in several of the main types of academic and professional careers. Manipulative-mechanical ability, mathematical ability, scientific ability, leadership ability, and ability in drawing, painting and sculpture are illustrations.

2. A more serious effort should be made to devise non-verbal tests capable of bringing out differences in general intelligence on the higher levels. The need for such scales both for individual and group examining is very urgent. If serial tests were used it would probably take about three such scales to cover the range from five-year ability to the higher adult levels.

3. The time is ripe for the development of an "infant" scale for measuring intelligence on the levels below three years. More reliable methods of diagnosis at this period would save a certain proportion of individuals from mental deficiency.

4. The Binet scale needs to be thoroughly overhauled in the light of recent progress in test methods. The scale should be made over in at least two or three forms, interchangeable and non-duplicative. Each should require not more than thirty to sixty minutes for its administration and should be considerably more accurate as well as more convenient to give and score than the present Stanford Revision. The writer has already undertaken this task, which will of course require several years for its completion.

5. Investigations are needed, of a more intensive kind than have yet been made, of a large number of individual tests for the purpose of determining their best possible composition and form, the best procedure for giving and scoring them, and their exact relationships to a variety of other tests. Any one of at least forty or fifty tests which might be named would each amply justify two or three years' work of a competent graduate student.

6. Next steps in the application of intelligence tests to scientific and practical problems include investigations of race differences, mental inheritance, the psychology of genius, normal and abnormal

mental growth, environmental influences, mental fatigue, and the relation of general intelligence to moral traits, social adaptability, and success in school and the vocations.

7. Investigation of instinctive, emotional, and volitional traits and of the combinations of these which are involved in pre-psychopathic conditions and normal variations in temperament. As this type of investigation is largely outside the scope of the intelligence problem its importance need not be dwelt upon here.

8. Early accomplishment of certain of these next steps is contingent upon finding money to finance more extensive investigations than we have yet become accustomed to. Thoroughgoing work in this field is extremely costly. My experience in helping to spend \$25,000 in the manufacture of one intelligence scale has convinced me that nothing less than \$100,000 is really adequate for such a task. The sum of \$1,000,000 is needed for a ten-year program of systematic, organized research looking toward the manufacture and standardization of eight or ten scales for measuring general mental ability. As much more is needed for research in the measurement of special aptitudes.

III. By F. N. FREEMAN,
University of Chicago.

1. *The nature and measurement of intelligence.* I conceive intelligence to be a somewhat more inclusive capacity than is implied when it is used as a name for our present tests. For this reason, it seems to me that it would be better to use a term of somewhat narrower connotation to designate these tests. The mental capacity designated by the term intelligence seems to me to include, besides the elements which are usually measured by our tests, certain other types of capacity which they measure scarcely at all.

The capacities measured by our tests may be described from the structural point of view as involving chiefly: sensory capacity; capacity for perceptual recognition; quickness, range or flexibility of association; facility in imagination; span or steadiness of attention; quickness or alertness in response. Excellence in capacities of this sort is well designated by the term brightness. We recognize, however, that a person's achievement, or even his intellectual capac-

ity, does not seem to correspond perfectly with his capacity in any of these various specific traits or with all of them put together. We sometimes say that a person is very bright but not very intelligent. We mean that he does not use his mental powers in such a way as to be most productive. The characteristic which I am referring to is sometimes called temperament or moral character, but it seems to me it can be shown that there are certain intellectual traits which are left out of our present scheme of tests and which should be included in a total conception of intelligence.

These additional characteristics may be described as: mental balance; co-ordination of the mental processes; the judicious management of the processes of learning or reflection; mental control; mental adjustment; the direction of the attention toward the significant aspects of experience; a due degree of non-suggestibility; the adoption of intellectual purposes and the adaptation of means to their satisfaction; sensitiveness to significant combinations between experiences which illuminate one another or which are effective in building up systems of thought; balanced and sane reaction to the entire world of things, ideas and persons.

2. *Next steps in research.* The attempt to catalogue some of the characteristics which would ordinarily be recognized as belonging to high intelligence indicates, it seems to me, that our tests are much too fragmentary to be thought of as even approximately complete measures of intelligence as a whole. There is undoubtedly a correlation between brightness and these broader aspects of intelligence, but we ought to devise tests to measure such processes as these more directly than is done by our present tests. This is one of the "next steps."

The second step is to devise tests of more specific capacities. The present-day general tests arose from the discovery that better correlation could be obtained between tests and other measures of undifferentiated ability, such as school marks and judgments, than could be obtained with simpler tests. A group of tests, each one of which has a low correlation with school marks, will give a considerably higher correlation when their scores are combined. There are various theories to explain this fact. It is not entirely clear whether such a combined score gives merely the composite result of a number of different abilities, or whether it is the measure of a central ability which is common to the various particular reactions and emerges

as the common factor. Whatever may be the cause, the practical bearing of the discovery is clear.

While these composite tests are of more practical value than those which give no correlation whatever with measures of every day achievement, they are far from satisfying our practical demands. Whether the pupil's intellectual capacity is a combination of general ability and of unrelated specialized abilities, or a combination of various abilities which are related to one another in varying degrees, the practical treatment of the individual requires that we should be able to measure his relative ability in more or less specialized types of activity.

This demand for tests of more particular capacities may appear reactionary to one who is acquainted with the history of mental tests. It was with this aim that the earlier tests were designed, and in seeking which they failed. It is obvious therefore that our procedure must differ from the early procedure if we are to succeed. There are two ways in which we may conceivably improve upon the early efforts. In the first place, we may utilize the improved technique with which the material is selected and organized, with which the test is presented to the child, the responses scored, and norms established.

There is another respect, however, in which it is necessary to improve upon the earlier attempts to test specific capacities. This consists in a different and improved definition of the components of ability. The facts of correlation present many puzzles from the point of view of our ordinary classification of mental abilities. It is common to find, for example, that abilities which are ordinarily classed under separate heads, according to our usual psychological system, correlate more closely with one another than do abilities which fall within the same field. One explanation which is offered for this fact is that the abilities which correlate with one another possess the hypothetical common factor in high degree. Another possible explanation is that there are mental traits or abilities which run across our ordinary classification. If we could discover what these abilities are, we might be able to determine and measure groups of abilities which correlate closely with one another but not so closely with other groups. I am inclined to think that there is one example in which we would find that our new classification of abilities would agree with our customary psychological system. This example is

memory. It is of course quite probable, and in fact almost certain, that an individual's memory for various types of experience differs considerably. On the other hand, it seems to me to be a hypothesis worthy of investigation that the memory or retention of experience is a characteristic of an individual's mental life which is rather distinctive and which is a more or less constant factor in his various mental processes. At any rate, no systematic experiment has been made to determine whether this is true.

Two important "next steps" appear to me, therefore, to be to measure broader aspects of intelligence, and to devise specific measures of significant components of ability.

IV. By S. S. COLVIN,

Brown University.

1. *Nature and measurement of general intelligence.* General intelligence has been defined as "general mental adaptability to new problems and conditions of life." To my mind this definition is somewhat too narrow. In a very true sense intelligence is mental adaptability to environment. This conception, however, is in one respect too broad since it includes instinctive adaptations as well as those that have been acquired through experience. Of course, psychologists sometimes speak of the psychic life of micro-organisms and frequently use the term intelligence in connection with instinctive acts of such animals as ants, bees and wasps, whose adaptations to environment seem to be almost entirely on the plane of instinct. On the whole, I consider the most helpful viewpoint from which to consider intelligence is that it is equivalent to the capacity to learn. *An individual possesses intelligence in so far as he has learned, or can learn to adjust himself to his environment.* In a sense this conception is substantially the same as that quoted in the definition first given. However, it does not unduly emphasize the problem aspect of intelligence and rightfully attributes intelligence to those animals whose sole ability to learn is confined to the hit-and-miss try-out of experience ("trial and error").

Psychologists have accepted this definition practically if not theoretically. An inspection of intelligence examinations clearly shows

that those who framed them have not confined their tests to problem solving, even in its rudimentary forms. These tests measure an individual's intelligence largely in terms of what he has learned, thus obtaining indirectly a measure of his learning ability. Vocabulary tests, range of information tests, "same and opposites" tests, tests of fundamental operations in arithmetic, and the like, call for little ingenuity. If the individual has the requisite skill and knowledge he can satisfactorily perform these tests. They are appropriate tests for intelligence only on the theory that they test ability to learn by discovering what has already been learned. Even those tests that demand verbal and mechanical ingenuity are valid only in so far as individuals taking these tests have had common opportunities to learn the elements necessary in solving the problems involved in sentence completion, thought interpretation and the like. It must be remembered that even the ability to think in a sustained and logical manner is based on having learned how to think. Thought is a habit and is acquired through learning. In a word, the validity of all mental testing rests on the fundamental assumption that those tested have had a common opportunity to learn the skills, facts, principles and methods of procedure exemplified in the tests. It follows for this reason that many of the standard school tests now in use are reasonably good measures not only of specific aspects of acquired intelligence, but also of general (innate) intelligence. All the individuals in the school group tested have had the same training, or at least very similar training. Some have learned more, others, less; those who have learned less possess less learning capacity,—hence less general intelligence.

Since, however, general intelligence cannot be considered as a single unitary factor (according to the Spearman-Hart-Burt hypothesis), but as a common average of many different factors positively, but by no means perfectly correlated, intelligence tests should explore as many aspects of human ability as possible. This is important for prognostic purposes. It is even more important if intelligence tests are to be employed to diagnose varieties of mental abilities. The investigator frequently finds individuals who do well in certain kinds of mental tests and who do poorly in others, not because of difference in opportunity, experience or interest but because of difference in native ability. We need the simpler tests, tests involving specific knowledge of facts, memory, perception and the like,

but we need also, and in a greater degree, tests to measure the higher intellectual processes,—tests that will give the individual who thinks carefully, accurately, but sometimes ponderously, an opportunity to show his ability. Doubtless speed of learning and efficiency of learning are positively, but by no means perfectly, correlated.

2. *Next steps in research.* As I have already pointed out in the previous discussion, we need at present test elements that emphasize, more than any now existing do, deliberation and sustained rational ability,—tests in which speed is relatively unimportant and in which analysis, synthesis and an extensive attention-span are the chief factors of importance. Pioneer work in this field, as indeed in many others, has already been done by Thorndike, particularly in his tests for college freshmen.

In one sense of the word there are too many mental tests at present. This plethora is doubtless valuable and necessary, as far as theory of mental testing is concerned, but it has definite practical drawbacks. The tests now “on the market” (and doubtless their commercial value has had something to do with their recent rapid development), while in general valuable, have too many features in common and are too nearly of equal value for practical purposes to make them all necessary. I hope the time will soon come when a committee of skilled psychologists will select the elements most valuable in the tests now existing, add others that are lacking, and after carefully standardizing this complete test, will issue it as the one recommended for general use in the grades and for the ages for which it has been devised. Of course, there would still be several tests,—one for the primary grades, another for the intermediate and grammar grades, one for the high school and one for the college,—but there would not be a multiplicity of tests for each level of school development, and there would be definite norms established for the guidance of teachers. At present either norms are lacking or they have been imperfectly and inadequately devised. And, by the way, would it not be well in arriving at standards to check back the results of the tests on groups of known intelligence and ability?

A further step that is necessary from the standpoint of the practical value of mental tests is that teachers and administrators should be more carefully informed as to the value and limitations of the results of intelligence testing in solving problems of instruction and supervision. Frequently tests are given and the results are in no

way utilized. Often tests are given and the results are wrongly interpreted and applied.

The most important "next step" for purposes both of prognosis and diagnosis is the formulation of a test that will inform us of the character qualities of those tested. It is true that there is a positive correlation between the results of intelligence tests and character, partly because intelligence and character are related and partly because our so-called intelligence tests are to an extent character tests as well. However, there are many instances in which intelligence tests fail to be of value practically because they give only slight indication of those qualities of character and temperament that are vital in all human achievement. In my work with students at Brown University, I have found scores of instances in which intelligence tests have not only failed to indicate in a positive way college performance, but have also shown results at variance with this performance. In a considerable number of instances the lack of relation has been clearly due to the fact that qualities other than intelligence have played a deciding role. The psychologist who devises a character test that has a reasonably high validity will earn for himself a position in the field of ability testing equal to that of Binet,—yes, even higher, for his problem is more complicated and his task more difficult. However, until such character tests are available we shall have solved only one-half of our problem in the prognosis and diagnosis of those elements which lie at the basis of human achievement.

V. By RUDOLF PINTNER,
Ohio State University.

1. *Nature and measurement of intelligence.* I have always thought of intelligence as the ability of the individual to adapt himself adequately to relatively new situations in life. It seems to include the capacity for getting along well in all sorts of situations. This implies ease and rapidity in making adjustments and, hence, ease in breaking old habits and in forming new ones. Fundamentally, this leads us back to the general modifiability of the nervous system. An organism whose nervous system is very modifi-

adapta

able can adjust itself quickly to new types of situations and in this way react with more intelligence to a greater number of situations.

The reactions to which we apply the terms "intelligent" or "unintelligent" are differentiated from those to which we apply the term "emotional," although any sharp division between these two types can hardly be made. Nevertheless, the distinction is a practical one which it is useful to retain.

Binet laid the emphasis upon judgment and reasoning, and in doing so turned the trend of investigation from the simpler to the more complex mental processes. Before the work of Binet, psychologists had been mainly concerned with the measurement of the simpler processes, and their work had not resulted in anything very practical from the standpoint of the clinical psychologist or educator. It was primarily the work of Binet that stimulated the use of tests of the more complex mental processes and led us to the combination of complex tests with which we are familiar today.

In no field of investigation is dogmatic assertion to be tolerated and certainly not in this new and vaguely defined field of mental testing. Unquestionably mental tests calling into play the synthetic and analytical activities of the mind, reasoning, judgment and the like, have up to the present time been most useful for practical measurement purposes. They seem to reproduce the complex type of situations common in modern civilization. We dare not, however, dogmatically assert that these tests are the only ones that ever will yield satisfactory measures. It may be that the simpler and more elementary processes, upon which the more complex depend, might prove useful for the measurement of intelligence. Up to the present time no one has been able satisfactorily to use them, but there would appear to be room for further research in that direction, even though the field at present seems rather barren.

Because we are dealing with something which we hardly know how to define, because we are groping for means to measure reactions to extremely complex and different situations, we must be very careful not to limit in any way the type of material we use. Intelligence is shown in dealing with things as well as with words, in dealing with living men and women as well as with symbols, in handling a paint brush or chisel as well as a pencil, in carving stone or laying brick as well as in making verbal responses. We, therefore, need intelligence tests of all types. And by intelligence tests I mean tests of

the general ability to do all sorts of things as opposed to educational and trade tests which are specifically made to measure the knowledge which an individual has been directly taught. If we believe in this large view of intelligence, we are painfully aware of the paucity of tests that exist at the present time. I cannot believe that we have exhausted our ingenuity in the construction of different types of tests and I hope to see many new types of tests constructed in the future. All sorts of tests and various combinations of all sorts of tests are needed.

As a matter of fact when we are confronted with the practical situation of planning a school survey, there are not many tests from which to choose. There are very suitable tests for the first and second grades. There are a few more available for the other grades. There are all too few really adequate tests for the high school and most of them are surprisingly alike. Really good discriminative scales for college students are rare. When in addition we look for scales with adequate standardization we still further decrease our list. So there is plenty of room yet for tests, for different kinds of tests and for well constructed and standardized tests.

Occasionally we hear the cry raised that of the making of new tests there is no end and that what we need most now is to refine our instruments of measurement, equate them, and standardize them. To refine, equate, and standardize would be good, but merely to do that would mean stagnation. The perfect measuring instrument of intelligence is not yet at hand and we need all the attempts, however crude and bungling some may be, in order to help achieve our aim.

Most of us can remember the horror of the "pure" psychologists at the advent of the Binet-Simon Scale and the horror grew into dismay as, in the course of time, article upon article poured from the press and threatened to turn all psychological periodicals into Binet journals. We were counselled and advised with, we were told to stop, we were told to go back to essentials and first principles and what not. Even some of the Binet workers themselves were dismayed and tried to stop the flood and advised staying where we were, until we had refined and analyzed the Binet monster. It is good that we did not, even although we might have had a refined Binet scale better than the best we have at present. In the meantime, however, the flood rolled on and brought in group tests of all kinds and the wider outlook on mental testing which we have achieved at

the present time. And it is to be hoped that mental testing will go on expanding and broadening in the future.

2. *Next steps in research.* I have already indicated in the previous paragraph one of the next steps in mental testing, namely the construction of various types of tests for the measurement of the various aspects of intelligence. We must not allow ourselves to imagine that the type of group test most common at the present time is the only measure of general intelligence.

Another profitable line of work is to make the tests function in the school and in life in general. If we have obtained some reasonable measure of intelligence, the value of this is very significant, and I do not believe we have really begun to appreciate the uses to which it can be put. A combined use of various mental and educational tests, suggested by several workers, is opening up profitable lines of research both theoretical and practical. The systematic use of these different measures must lead us to raise important educational questions. The amount of educational attainment to be expected of any specific degree of intellectual ability can be stated in more definite terms than we have been able to state it before. The practical bearing of this upon the efficiency of the school must be carefully studied. Annual or semi-annual surveys of schools should be systematically undertaken. Various methods for the elimination of wasted intelligence should be introduced as a result of the findings of the surveys, and the value of each such method should be studied. Our question would be: To what extent does this or that method actually decrease the amount of wastage? In other words, I feel that the time has now come for systematic measurement of the same group of children over several years, by means of both educational and mental tests, and combinations of these; and along with this a study of educational devices for increasing efficiency.

So far I have been discussing mental tests in the narrowest sense of that term. I feel, however, that the time is now ripe for active investigation of the emotions, the character, the will and so forth, by means of mental test methods. A beginning in this work has already been made by a few workers. It is to be hoped that work in this field will increase rapidly, so that we may learn what possibilities exist for objective measurement of these aspects of the individual. Personally, I feel that the possibilities are great, and that in the near future we may arrive at some quantitative expression of

the emotional side of the individual. This will bring us one step nearer to our ideal, namely, a psychological profile or equation of the whole man.

VI. By B. RUMML,

Carnegie Corporation, New York City.

1. *Intelligence and its measurement.* It seems to me that the question as to the nature of intelligence can hardly be debated at the present time. There are two reasons for this: first, the lack of precision in the terms and concepts that must form the basis for such a discussion and second, the absence of factual material on so many of the essential points. Much of the best data that is available is compromised by inadequate sampling and controls, and by contradictory reports from various observers. For example, the status of the "common factor", "general intelligence", is extremely unsatisfactory; and surely the general problem of intelligence requires for its solution an accurate statement concerning this alleged "common factor".

As to the best kinds of mental tests, I feel that test materials of every kind are still desirable in test research. We need data concerning the simpler mental processes as well as the higher. The one limitation that may properly be made is that in any test which yields a single quantitative result or score, the content should be as homogeneous as possible. Valuable as omnibus tests and mixed test series may be for practical purposes, it seems unlikely that these will assist in any really important way in the understanding of intelligence.

2. *Next steps in research.* The most promising "next steps" will perhaps come through the study of changes in test score and in inter-test correlations with change in chronological age.

For example, for certain tests there is reason to believe that in spite of considerable variability, there is practically no correlation with chronological age between the years of 9 and 13. How does this come about? Will it not be illuminating to know more completely than we do now what types of tests behave in this manner? It is, of course, clear that we are dealing with relationships that

show non-linear regressions, and the commoner statistical approach will probably be inadequate.

Again, there is some slight evidence that inter-test correlations are lower for children of a given age than for adults. We have little data of a quantitative kind to show how great this difference is, neither do we know how these relationships behave for various kinds of test material. But if with more factual material the reality of this difference in size of inter-test correlations is established, the attempts at explanation will be sure to develop interesting and important light on the general problem of intelligence. Research of this type requires very accurate sampling of subjects and is, of course, very difficult on a small scale.

A third attack of some promise is suggested by the "will-profile" technique of Dr. Downey. The methods used by Dr. Downey in constructing her test materials are rather different from the methods of building many intelligence tests. They suggest the need for a much more exact qualitative study of the test situation than is ordinarily made as a preliminary to final standardization in quantitative form.

Finally, there is need for more adequate methods of representing the results of test performances. The profile gives a good deal of information, but it is difficult to interpret, and is rarely used in providing data for generalization.

VII. By S. L. PRESSEY,

University of Indiana.

1. *Intelligence and its measurement.* The writer has been asked for a statement as to what he conceives intelligence to be, and as to what types of test materials can best be used for the measurement of intelligence. Frankly, he is not very much interested in the question—although a large part of his time goes to work with tests of "intelligence". Instead, he is interested to know what such tests will do, in solving this or that problem. Suppose, for instance, one is intending to survey a school system in order to find those children who should be put into special classes—classes for those so dull as to be incapable of profiting by the regular school work. Tests in the

school subjects cannot be used for this purpose, since some children would then test low because they had had their first training in poor country schools, or had been out of school because of poor health. Under these circumstances, one naturally makes up an examination not involving matters taught specifically in school, but a test as much like school work in other respects as possible. We have, then, a "general" test.¹

After a number of such tests have been made up the next step is, naturally, to evaluate them according to the efficiency with which they deal with the practical problem for the solution of which they were made. The obvious procedure is to give the tests to (1) a group of children who are known, from long and intensive study of them, to belong in special classes, and (2) another group known to belong in the regular classes. The tests which separate these two groups most distinctly are, then, the tests to include in the final scale.

Measurement of capacity for school work by such a paraphrase of school work, but with the specific elements of schooling left out, is surely a very natural thing to do. It is also natural to attempt the selection of children at the other extreme—the exceptionally bright children, who are capable of more advanced work than they are now doing—by means of such tests. The capacity of these children for this advanced work cannot be determined by giving them some of this advanced work to do; the point of the whole situation is that opportunity has not been given to acquire the specific elements in this advanced work. In fact, *any* effort to get back of the school's actual disposition of its children, into classes and grades, naturally suggests the use of such paraphrase of school work, but with the specific elements eliminated. And the usefulness of different types of this material must be determined by actual trial with the concrete problem to be dealt with.

Well—and if such tests are of service in dealing with a variety of problems one may come to think of them as tests of a fundamental, unitary, general intelligence. An elaborate theory, with elaborate statistical expression, may then be developed. But the hypothetical nature of the concept must be kept in mind (and the educational situation which first gave rise to such tests, also). Particularly,

¹The problem described is, of course, the problem which Binet faced in 1905 and which gave rise to the original Binet Scale.

one must not permit himself to be hypnotized by the statistics. Always, it must be kept in mind that there is no value in measuring intelligence simply for the sake of measuring it; the tests are given in an effort to deal with some practical problem. And it is in proportion as they deal efficiently with that problem that they are to be considered satisfactory.³

Frankly, then, the concept of general intelligence is for the writer, primarily, simply a working hypothesis which has been very helpful in the first attack upon problems of prognosis. It should continue as a working hypothesis until other hypotheses are found to work better. But investigators must not be so dominated by the concept of general intelligence that they fail seriously to consider other possibilities. Particularly to be avoided is the easy notion that a test of intelligence will solve all educational difficulties, that intelligence testing is the longed for short-cut method which will put everything in the schools finally and completely to rights. The theory is simply an assumption that prognosis problems can be generalized. It has been of great use in suggesting certain general lines of approach,—and in giving us courage to attack the otherwise impossibly complex prognosis problem. It will do equally great harm if it prevents analysis of the prognosis problem, now that methods are becoming sufficiently refined to make this possible. The safest, most empirical, and most practical method of making this analysis (to the writer's mind) is to study the relation of the tests to particular practical problems. From such study any general factors will, in time, emerge of themselves.

2. *Next steps in research.* Effort in two directions is especially needed; in neither case should the study deal with general intelligence:

(1) It should be obvious that even if certain abilities (as in arithmetic) were highly specific, this would never be discovered if scores on tests were always summed into a single rating in "general intelligence", marks always averaged, or teachers asked to estimate general ability. Of course, under such circumstances, there will appear

³The most fundamental statistical problem of all must surely always be: "Are these methods applicable to the data and problem in hand?" Instead, there seems to be a tendency to regard a thing as so, if it can only be squeezed into a formula. And we find ourselves, before long, talking easily about absolute point scales, zero points in intelligence, and so on—as though we knew all about it. Such practice might, perhaps, be called "statistical transcendentalism."

⁴The writer has presented this point of view more fully in a recent paper, "Suggestions Looking Toward a Fundamental Revision of Current Statistical Procedure, as Applied to Tests." *Psychological Review*. Vol. 27, pp. 466-472. November, 1920.

to be a general factor, because everything has been generalized! Yet the results of such procedures are used over and over again as evidence of the all important character of innate, general ability. The situation verges on the absurd. There must, then, be intensive critical study as to the concept of general ability—and a more open mind on the subject than seems to be common at present.

(2) In the second place, there simply must be a courageous attack upon the problem of measurement of other than intellectual factors. It is becoming increasingly obvious that matters of temperament and character are of very great importance, that they operate quite largely independent of intelligence, that prognosis problems cannot be adequately understood without an evaluation of these factors. It is also probable that these factors are more educable (in the large meaning of that word) than intellectual traits. From every point of view, then, work in this field seems most desirable. Some beginnings have already been made. It is in this direction especially, the writer feels, that research effort should be directed for the next few years.

THE GROWTH OF INTELLIGENCE AND THE INTELLIGENCE QUOTIENT

JOSEPH PETERSON,
George Peabody College for Teachers.

Though the measurement of intelligence has of late made notable advances in its practical applications, we are yet considerably in the dark when we ask detailed questions regarding the interpretation of our results, questions which are not mere academic puzzles but which have a number of important practical bearings. That this is the case is evident when we attempt to interpret in a comparative way the norms now available for the different kinds of mental tests. In a recent article in this journal¹ Freeman has attempted to call in question certain assumptions that seem necessary for the validity of the IQ as used in Binet testing. The assumption of "divergence of growth curves as well as a decrease in rate of growth," made by Woodrow in his *Brightness and Dullness in Children* respecting the development of intelligence is questioned; and, after pointing out that the yearly increments in certain of our group intelligence norms are about constant through a number of years, Freeman concludes as follows: "It appears from these facts that both the assumptions which may serve to explain the validity of the IQ in the case of the Binet Scale are in question." (P. 12.) The validity of the IQ for prediction depends, of course, on its constancy throughout a number of successive years; and since the IQ is a ratio of mental age to chronological age, it is obvious that the lines of intelligence growth of individuals of different IQ's must diverge, so that a retardation of one year at an early age will equal in effect one of two or more years in later stages of growth. Moreover, speaking in general of the rate of growth of intelligence, "It may be shown mathematically that if the rate of growth were the only factors determining the IQ, and if the IQ were valid, the curve would be logarithmic, the formula

¹Freeman, F. N. The Interpretation and Application of the Intelligence Quotient. *J. Ed. Psychol.*, 1921, 12, 3-13.

being $y = \log x$.”² Freeman produces the curves of a number of point-scale norms to show that the logarithmic form of the growth of intelligence is not in evidence. Most of these curves are in fact nearly straight lines, some of them being exactly straight, through several years,—that for Yerkes’ point-scale tests being straight from years 4 to 12; for Pressey’s, from years 8 to 16, with a slight divergence covering years 11 to 14; and similarly for the Haggerty tests, the Pintner non-language tests, the National Intelligence Test, and others.

From these “curves” Freeman seems to read off directly the normal rate of the growth of intelligence; that is to say, he seems to regard these curves of norms for point-scale tests as intelligence-growth curves, in as much at least as these tests actually measure intelligence. While he admits that these curves offer certain difficulties—such as a gradual bend toward the X-axis in certain cases at about the beginning of adolescence, possibly due to the nature of the tests themselves, and also an occasional small break in the curve—and that the point of actual slowing up of growth in intelligence can be located only after more extensive investigation, he concludes that “The preponderance of evidence, however, seems to indicate that up to some age in early adolescence at least the rate of growth is approximately uniform.” (P. 9.)

This is precisely the point, it seems to me, at which Freeman’s whole position must itself be called seriously to question. His position seems to be founded on the view that these curves are in themselves intelligence-growth curves. Except on the most superficial view of them, one that unfortunately is too often taken, this position is wholly untenable. The curves, indeed, show only the *number of points scored* at the ages indicated under certain standard conditions, the most important of which is a definite limitation of time, constant for all ages.

In the first place, it is hardly justifiable to apply to the Binet scores from which the IQ is obtained, inferences of this nature from scores derived under this time-limitation. In the Binet tests there is practically no time limit in most cases, and the tests vary from age to age with considerable overlapping. This is, of course, why we

²Freeman, *op. cit.*, p. 5. It is obvious, however, that the simple logarithmic curve does not hold strictly, especially at the extremes of age, for the growth curve of intelligence. There is some degree of intelligence, probably a comparatively large amount, in the year-old child, and there is also rather good evidence that its growth does not continue to senility, whatever we may conceive intelligence to be.

have had from these tests no very serious attempts to establish any well defined view regarding the exact nature of the increments of intelligence development through successive years, though the logarithmic curve has been implied as a rule as indicative of the rate of mental growth.

Secondly, and this is our most important point, making inferences as to the nature and rate of intelligence growth—either relatively to earlier periods or absolutely—from absolute scores of average accomplishment in definite periods of time, is a business that is beset with hazards for a young science, to say the least. A concrete illustration will tend to bring out the difficulty in question. The 1920 norms of the Otis Scale, Advanced Examination, are as follows:

TABLE I.

	Age.										
	8	9	10	11	12	13	14	15	16	17	18
Total time allowed.....	42	42	42	42	42	42	42	42	42	42	42
Standard score.....	40	52	64	76	88	100	112	121	125	128	130
Increase in score in one year...		12	12	12	12	12	12	9	4	3	2

It is obvious that the equation

$$W = ta \quad (1)$$

will hold, if W represents the work done, while t and a represent, respectively, time, or number of minutes, and the average amount of work done per minute. Applying this equation to our special case, and letting W stand for the total number of points scored and a for the average number per minute, we may for the present purpose disregard any "work" that does not make toward the production of scores. We are compelled here, moreover, and in what follows, to assume that each score-point has equal value with every other score-point. Now figuring the values of a for the total scores given in Table I for years 8 to 14, through which the scores increase uniformly so far as the absolute units are concerned, we get the values for a indicated in Table II. It is to be noted that the time of 42

TABLE II.

	Age.						
	8	9	10	11	12	13	14
Value of a95	1.24	1.52	1.81	2.10	2.38	2.67

minutes for the Otis test is exclusive of time used for instructions and all preliminaries; it is the time actually devoted to the work of score-making.

Now it is obvious that a also makes a constant increase from year to year, the increment being about .286, whereas the increment in the total score is 12, as shown in the first table. Equation (1), when applied in the present case to the changes in ability through successive years, may be put into the form

$$W + nK = t(a + nk), \quad (2)$$

in which n indicates the number of years after the eighth, and k and K represent, respectively, the constant yearly increments in average per minute and in total score. In these equations it is seen that a and t hold inverse relations, speaking roughly, and that therefore making t constant, as is done in the point-scale group tests, makes the average-per-minute values vary directly with the total score. Is any one going to maintain seriously that an increase from .95 of a score-point per minute, at eight years, to 1.24, at nine, (31%) is indicative of the same growth in intelligence that an increase from 2.38 to 2.67 score-points (12%) from year thirteen to year fourteen represents? The ratio of increase decreases perceptibly from year to year. Shall we base our view of mental growth on the absolute increments or on the proportionate decrements in time for doing a specified amount of work? Probably no one knows. One's choice will depend on just what one means by growth of intelligence from year to year and in what relationships one is considering such growth.

The importance of this point comes out clearly if we ask ourselves: How does the speed decrement, *in absolute units*, for doing a constant amount of work—making a given score-point—behave through successive years? From the data given in Table II, we can determine the average number of minutes per score-point for the successive ages there shown, or the values of t as given in equation (1) for successive years of mental growth, if W is made unity. Thus we find in Table III the number of minutes required by children of the successive ages shown, to make on the average a score of one point. Figure I shows both the absolute increments of a and the absolute decrements of t as given in the tables.

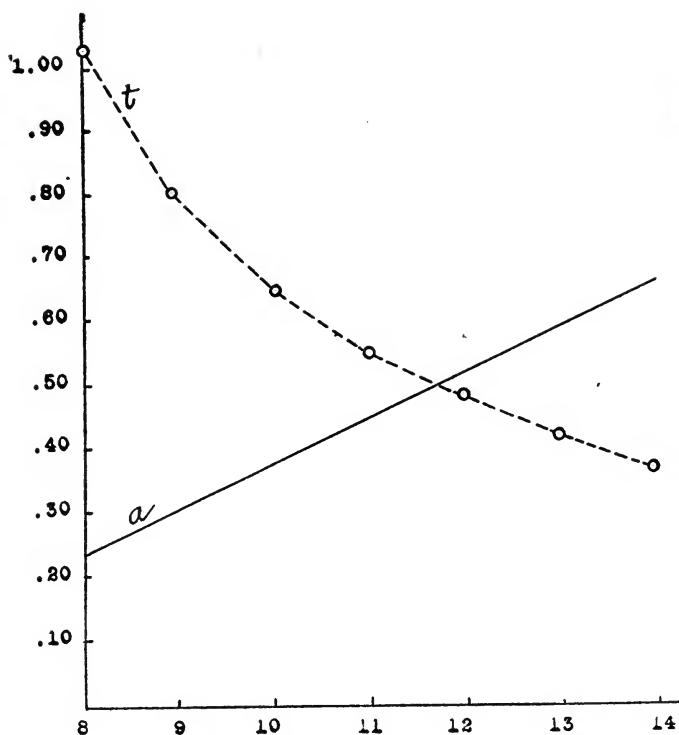


FIGURE 1.

The curves are norms of the Otis Scale, Advanced Examination. Curve *a* shows the average number of score points made per minute, while curve *t* indicates the average number of minutes required for each score point for different ages. The ages are shown on the base line.

TABLE III.

	Age.						
	8	9	10	11	12	13	14
Average number of minutes to make one score point.....	1.05	.81	.65	.55	.48	.42	.37
Decrease over previous year...		.24	.16	.10	.07	.06	.05

From a consideration of the data of tables II and III and of Figure I it is obvious that one might come to different conclusions as to the constancy of the rate of mental growth as indicated by performance at different ages, according to whether one considers *how much work is done in a unit of time*, on the one hand, or *how much time is required for doing a definite unit of work*, on the other. It is to be

noted that in either case here one considers absolute increments or decrements. The conclusion from this consideration seems to be that we cannot read off mental growth directly from either of these score changes, as Freeman seems to have done from one of them.

The error here indicated is similar to one that has lurked in the inferences from learning curves as to the rate of learning, first pointed out by the writer, so far as he is aware, is 1917³, an error still persisting in certain recent articles on learning, even though the author in one case has explicitly noted our criticism and thinks he has avoided the error.⁴ The error in question here is also similar to one in the statistical determination of averages, which was pointed out by Rugg simultaneously with our own statement of it as applied to learning.⁵ In dealing with curves of complex functions, or with the data that such curves represent, it is essential that we do not neglect the possible effect of factors of importance that are not in evidence.

It is, of course, interesting to know of any test norms that the increment from year to year is about constant, and also to note when and how this increment begins to change with the approach of mental maturity; and such knowledge may be useful in many practical ways, even though applications such as we have questioned in this paper cannot be made in any direct manner. Such constancy of increase, however, does not necessarily indicate that a test is a good *mental* test, or even that it discriminates well the different degrees of intelligence known to exist in successive year-points in mental development. Moreover, other factors should be considered or controlled in interpreting such data. For example, let us suppose that the children become fatigued in inverse order to their ages—an assumption that is not unreasonable in long tests—and that therefore the attention and application in any test period constantly decrease, thus preventing the younger children from using profitably all their time. This will obviously lower their score beyond what more fair conditions would make it. Now if this effect appears early and gradually decreases with advancing years, it may exactly balance and therefore cover up the effects of real mental changes that we are attempting to measure; or it may tend to exag-

³Peterson, Jos. Experiments in Ball-tossing: The Significance of Learning Curves. *Jour. Exp. Psychol.*, 1917, 2, 178-224.

⁴Perrin, F. A. C. The Learning Curves of the Analogies and the Mirror Reading Tests. *Psychol. Rev.*, 1919, 26, 42-62.

⁵Rugg, H. O. *Statistical Methods Applied to Education*, 1917, 126-132.

gerate them. We are not trying to indulge in a hair-splitting exercise regarding this matter, but desire to show that a time element cannot be safely disregarded in the comparison of different norm curves taken as indicators of the rate of mental growth.

With reference to the problem before us, the successive yearly increments of mental growth in children, it is probably obvious that we can yet say but little with confidence, however useful the various intelligence tests have become in the applications of psychology. The article in question is far from making dogmatic conclusions, and its cautions regarding the applications of the IQ to other scales than the Binet, from which it originated, is timely; but if the criticisms here offered are well founded it is well to abstain from conclusions as to the absolute growth of intelligence, based only on norms of average attainment per unit of time. If the constancy of the IQ is supported by more extensive research than has been made we shall probably have to accept the view that mental growth curves diverge and also that they decrease in rate with advance of age, though it is very questionable that the logarithmic curve will be found to correspond strictly to these curves at their extremes.

COMMENTS ON PROFESSOR PETERSON'S CRITICISM.

F. N. FREEMAN,
University of Chicago.

The point which Professor Peterson makes with reference to the unit to be used to measure achievement in the age progress curve is worthy of consideration. In discussing it, we should distinguish clearly between two aspects of the question. My original paper dealt first with the use of the IQ as a measure of the mental capacity of the individual in connection with various scales, and, second, with the nature of intellectual progress as indicated by the scores obtained with these scales.

My argument concerning the IQ is not affected at all by Peterson's criticism, for the reason that it deals simply with the units which are actually used in these scales and not with the units which might have been used. Any person who uses the Otis scale, for example, will naturally score the pupils according to the units which Otis uses, namely, number of points; and the question whether the ratio of mental age to chronological age is a suitable measure has to be considered with reference to the type of age progress and distribution within each age which is found to exist in the measures which are actually used in scoring the test. For this reason, I did not, as Peterson implies, "apply to the Binet scores . . . inferences of this nature from scores derived under this time limitation."

Let us be clear on another point before proceeding further. My statement concerning the form of the progress curve might have been worded as follows and still have retained its chief significance: "The age progress curve *when expressed in terms of annual increments of point score* appears to be approximately a straight line." The significance lies in the comparison of such age progress lines with others, *all of which have been obtained with the same kind of units*. Whether or not what appears to be straight linearity by such units is really such, or is only something which more nearly approaches straight linearity than has been obtained from so many of the single tests—such, for example, as are reported by Pintner and Paterson in their "*Scale of Performance Tests*"—is a question chiefly of theoretical interest.

The theoretical question whether it is better to express growth in terms of absolute increments of score or relative decrements of time seems to me one which is open to debate. Peterson scouts the idea that an advance from .95 to 1.24 points per minute is indicative of the same growth as an advance from 2.38 to 2.67 points. Apparently the reason these two intervals seem so different is the fact that the lower scores are so much nearer the zero point than the upper ones. But we cannot take such an appearance, when we are dealing with data of this kind, at their face value. Suppose that the decrements shown in Peterson's Table II should be found to continue downward, as shown below. The series would reach the zero point

	Age.									
	5	6	7	8	9	10	11	12	13	14
Points per minute.	.09	.38	.66	.95	1.24	1.52	1.81	2.10	2.38	2.67
Minutes per point.	11.11	2.63	1.82	1.05	.81	.65	.55	.48	.42	.37

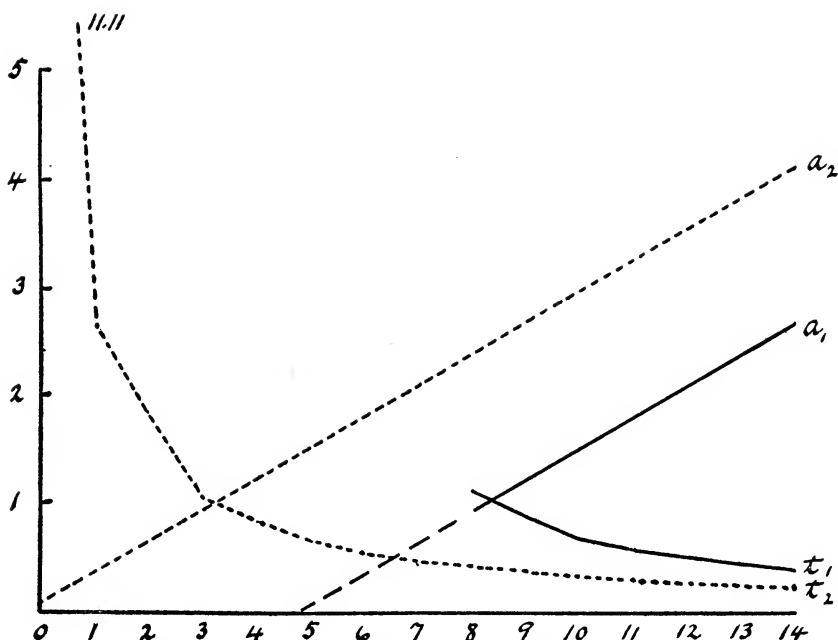
somewhere between ages 4 and 5. Obviously this is not the point of beginning of mental growth. The test is more difficult than it would have to be if a zero score represented zero ability.

The origin and the form of the time curve at given ages is affected by this difficulty level of the test. Thus in the present hypothetical extension of the curve downward the time curve reaches infinity between the ages of four and five.

Suppose that the whole level of scores were raised through the addition to the scale of easier tests enough to bring the zero score approximately to zero age with a continuance of the straight-line progress. This would be done by adding 1.43 points per minute to each score above age 5. Our series of scores would then run as follows:

	Age.														
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Points per minute09	.38	.66	.95	1.24	1.52	1.81	2.09	2.38	2.67	2.95	3.24	3.53	3.81	4.10
Increment29	.28	.29	.29	.28	.29	.28	.29	.29	.28	.29	.29	.28	.29
Minutes per point	11.11	2.63	1.82	1.05	.81	.65	.55	.48	.42	.37	.34	.31	.28	.26	.24
Decrement		8.48	.81	.67	.24	.16	.10	.07	.06	.05	.03	.03	.03	.02	.02

The numerical quantities are shown graphically in the figure. The form of the progress curve which is based on the time units for ages 9 to 14 is seen to be altered very seriously by the hypothetical readjustment of the level of scores to bring the zero point of the test more



a_1 , Norms for the Otis Scale in points per minute.

t_1 , Norms for the Otis Scale in minutes per point.

a_2 , Norms for the Otis Scale hypothetically raised to a higher base and extended downward in points per minute.

t_2 , Norms represented in a_2 expressed in minutes per point.

nearly where we should expect zero ability to exist, but the form of the curve based on points is not affected. This procedure, of course, is somewhat arbitrary, but it seems to me sufficiently valid to indicate that there is serious question of the validity of the time unit method.

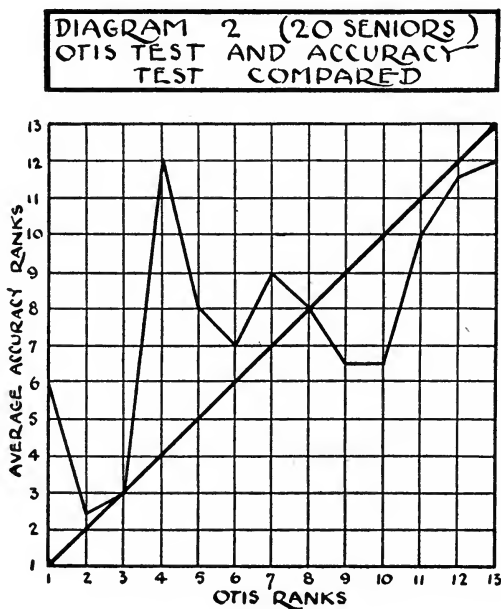
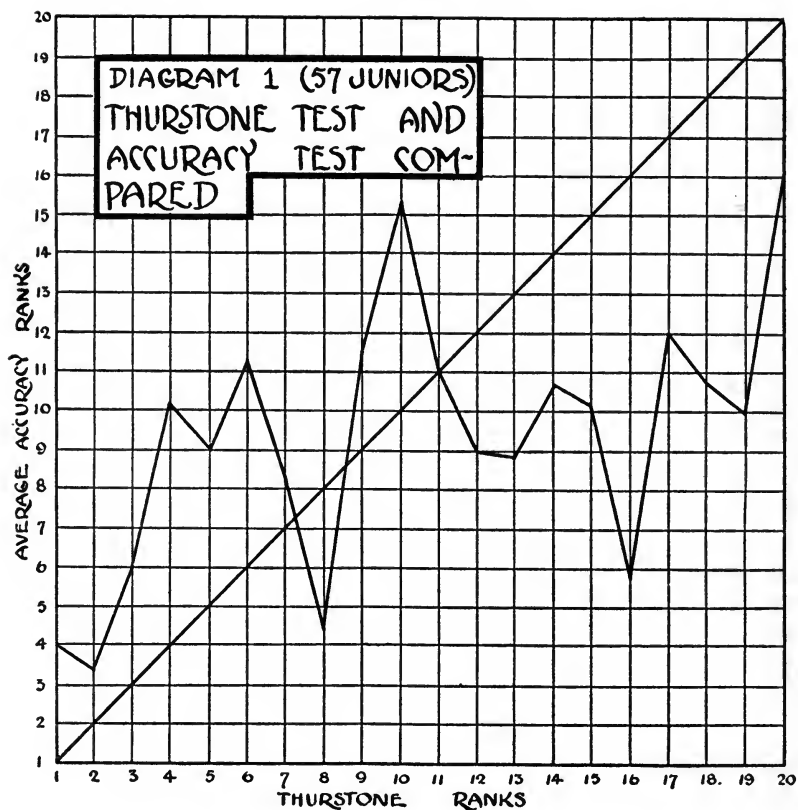
There is one other difficulty with the theoretical basis for the time unit method, and that is the assumption that tests of the sort we are discussing are purely speed tests. Most of them do, it is true, employ a time limit. But a still more important feature of their organization is the progressive increase in the difficulty of the parts of each test. The higher scores, therefore, represent not merely, or perhaps not chiefly, greater speed, but ability to do a more difficult task; and insofar as the higher score does represent greater speed the speed may be regarded as the outcome of greater power in performing the task rather than as greater quickness in itself. We are

accustomed to base growth curves of other sorts, such as height, on increments in amount rather than on proportionate increase, and it seems to me conceivable, at least, that mental growth may be at least as closely analogous to growth by increments of this sort as to increase in the rate of performance.

TIME AND ACCURACY AS RELATED TO MENTAL TESTS

E. LEIGH MUDGE,
Edinboro, Pa., State Normal School.

During the Autumn of 1920 the writer gave to the junior class of the Edinboro State Normal School the group test of general intelligence devised by Thurstone, and gave to the senior class the test prepared by Otis. A considerable number of inconsistencies between test scores and academic standing led him to wonder if certain valuable mental elements may not be disregarded by these and similar tests. Certain students who ranked relatively low in the tests seemed to do more accurate and careful school work than others whose rank was higher. Certain students seemed to feel peculiarly handicapped by the time limit imposed in giving the tests. Others, working at a much higher rate of speed, were able to make higher scores in spite of many inaccurate responses. Was it possible that the high scoring students were securing speed at too great a cost in accuracy? Granting the importance of quick responses in many or most situations, is not accurate work relatively undervalued in these tests? A series of nine group tests without time limit was devised for the study of these problems. These tests demand thought and carefulness but do not involve such a range of problems, from easy to difficult, as would be necessary for a thorough reasoning test. They are chiefly tests in accuracy in a variety of tasks involving reasoning, observation, memory, and other mental functions, especially accuracy in following directions. The only values which may legitimately be used in comparing the records of these tests are relative ranks in each test. Diagram 1 shows the relation between twenty rank-groups in the Thurstone test and the average rank of the same groups in the accuracy tests. The horizontal measurement indicates the Thurstone ranks, from left to right. The altitude of the curve indicates average rank in the accuracy tests. For example, those who are in rank 12 in the Thurstone test are of the average rank of 9 in the accuracy tests. The diagonal straight line is the line which the curve would follow if there were a perfect correlation between the rankings of the two tests. Diagram 2 follows the same plan, save that there are 13 instead of 20 ranks.



The striking similarity between these two curves appears to indicate, at least for this institution, some significant tendencies. Granting that the Otis and Thurstone tests are dependable measures of general brightness, and that accuracy of work, at least under the conditions described above, is measurably indicated by the tests devised by the writer, the following tendencies are observable:

1. The very few brightest students tend to be the most accurate, and the very few dullest students tend to be the most inaccurate.

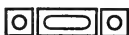
2. Aside from these extremes, the brighter students tend to be relatively *less* accurate and the duller students to be relatively *more* accurate. The reason for this may lie in an educational fault. Is there not a tendency to accept relatively careless work from the brilliant, rapid-working student, and does not the student himself feel that more is to be attained by working rapidly though carelessly than by spending more time at a given task?

3. There are persons who are relatively inefficient in such timed intelligence tests as we have used, who are able to do relatively efficient and accurate work when allowed to work more slowly.

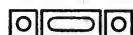
4. There are persons who are relatively efficient in rapid work, who are relatively careless and inaccurate, even when they may work more slowly.

The educational bearing of these observations seems clear. Since accuracy in observation and work is highly desirable, the schools should provide such work for the slower group as will develop the habits of careful work through which they may render their best service. The schools should likewise seek to so adapt their work to the brighter students as to more effectively motivate accuracy. Beyond a doubt, society needs people who think and act quickly, but their usefulness may be increased as they develop habits of greater accuracy and carefulness in their work.

DEPARTMENT FOR DISCUSSION OF RESEARCH PROBLEMS



Conducted by LAURA ZIRBES



This department has a two-fold function. It aims to serve research workers as well as educators, whose work brings them in close contact with children in the schools. It hopes to accomplish this service by suggesting research studies, which will meet well-defined school needs.

In order that this service may be real and effective, the co-operation of research workers and school people is desired. Correspondence with reference to the following questions will be considered in selecting topics for future discussions.

- a. Which of the studies proposed would help you to solve a practical problem?
- b. What topics might well be added to this list? Replies may be addressed to: Miss Laura Zirbes, 646 Park Ave., New York City.

1. *Studies in the psychology of practice which does not have an automatic response to a particular situation as its end.* The effect of mere recall without reorganization of elements. Experimentation to determine the value of diverse associations and reorganization of experience as a device to promote learning, recall, and transfer. Psychological differences in learning due to difference in purpose and content of instruction as illustrated by comparing psychological processes involved in learning addition combinations or mathematical formulae with learning the meaning of historical movements or geographical influences.

2. *A study of the effects of specific training in the appreciation of quantity and quantitative thinking on the ability to solve quantitative problems.* A study of the social value of the ability to realize and manipulate mentally number relations, followed by an experimental determination of socially valuable practice material.

3. *The effect of various types of motivation on performance in tests and practice materials.* Variation of individual scores due to desirable or undesirable attitude. Possible economy resulting from well motivated practice. Comparison of results of same motivation on children of different grades and ages. Evaluation of various types of motivation with reference to the particular needs of subject matter.

WHAT ARE THE SITUATIONS IN WHICH READING FUNCTIONS?

A scientific determination of procedure and content in the teaching of reading cannot proceed without a careful study of the situa-

tions in which reading functions. Present progressive practice as well as new reading textbooks show the influence of studies which have resulted in the re-appraisal of oral and silent reading. But a further analysis is necessary if instruction is to receive a full measure of the benefits which accrue from psychological studies and educational research.

Granting that the ideal curriculum represents the social values from the accrued experience of the race which the best judgment of this generation considers of such great importance, that they should be made available to all the members of the succeeding generation, does it not follow that the following questions can only be answered by careful research and investigation?

1. What do various types of American citizens read of their own choice? (What relation has this to the problems of recreation and citizenship?)

2. What does an American citizen need to read in order to pursue the round of activities which the intelligent performance of his duties as a citizen demands?

3. How can we, in our educational program, provide a sufficient variety of reading experiences and desirable types of training to demonstrate to learners the social significance of reading?

4. What are the special study habits and attitudes which make reading an effective tool in the pursuit of learning?

5. What aesthetic and recreational experiences depend on the cultivation of an interest in books and on the possession of means and ability to satisfy interests by reading?

Our times certainly make a wide range of demands on reading abilities. There is, perhaps, no school subject which compares with reading in making life significant. A great many people are able to evade the necessity of writing and making arithmetical calculations without seriously limiting their own efficiency or interfering with the enjoyment of leisure. But there are few indeed who can so easily evade reading even for a day, nor do they desire to do so. Individual reading abilities limit the amount read and are, in a large measure, responsible for kinds of material selected. While it might be interesting to study the actual reading habits of a random sample of our population, it would hardly be wise to use the results for determining the types of abilities which the school should try

to perfect. We should rather seek to provide every future citizen with a sufficient range of reading abilities to enable him to keep abreast of the times in the performance of his civic, economic and recreational responsibilities.

By an analysis of the reading content of papers, current magazines, and books, we find that there are many kinds of reading matter, each requiring special motivation and special abilities. The following list is by no means exhaustive: Easy narration, description, current news, articles, essays or editorials, correspondence, technical and scientific information, poetry, diagrams, charts and tables, summaries, jokes, serious business announcements, minute directions, etc. When various interests and abilities have not been developed it is not surprising to note that the corresponding materials are not read. Answers to the following questions would be illuminating:

1. What is the most widely read part of the Sunday newspapers?
2. What percent of the subscribers habitually read the editorials?
3. What percent of the population reads a foreign language by preference?
4. Who are the subscribers to the various types of magazines and how does the constituency vary?
5. What could be learned of the reading habits of various factors of our population by a study of the records of public libraries?

The mere statement of these problems leads to the formulation of others depending on the sociological definition of reading outcomes.

Silent reading has been stressed since its social significance became apparent. A great number of special habits depend largely on carefully designed training and experience. The mere pursuit of silent reading may lead to a very one-sided performance. We are beginning to realize that there are many silent reading abilities and that training to be economical and effective must be special. This can be easily demonstrated by tests.

Test material which consists of a very simple story to be read without interruption or reproduction must obviously be supplemented by other tests if a true measure of the pupil's ability in situations which involve reading is to be gained. If, as is apt to be the case, practice is redirected with reference to the requirements held up in tests, it follows that groups of tests should be designed to cover

the many socially desirable types of performance in which reading is a more or less dominant factor. This does not abrogate the usefulness of some of the carefully standardized tests, but it does indicate that there are gaps which need to be filled by the suggested tests.

This suggestion leads to another. After an experimental determination of the reading curriculum, the abilities which are involved must be analyzed psychologically to the end that the technique of instruction and practice with desirable content may be built up on a scientific basis. This, in turn, implies the necessity for diagnostic and remedial measures, for instructional practice material based on the psychology of reading tested and evaluated by the changes in ability which they produce under carefully controlled conditions.

Although some phases of reading have been very carefully and thoroughly investigated, further research along the lines suggested must make available to the profession scientifically established standards to be met by materials and methods of instruction.

If this discussion leads to the investigation of any of the problems suggested, this department will welcome an opportunity to outline more fully the field of investigation and possible ways and means of carrying out the suggestions offered.

LAURA ZIRBES.

The Lincoln School of Teachers College.

NOTES ON ARTICLES IN EDUCATIONAL PSYCHOLOGY IN CURRENT ISSUES OF OTHER MAGAZINES

1. ARTICLES ON TESTS.

An Experiment in Arranging High School Sections on the Basis of General Ability. Ernest P. Branson. *Journal of Educational Research.* January, 1921. 53-55.

Suggestions for Procedure Following a Testing Program. B. R. Buckingham. *Journal of Educational Research.* December, 1920. 787-801. Recommends re-classifying and working with individual pupils after giving group test. Gives the first bibliography we have seen on this problem. Bibliography deals with both educational and mental tests.

A Group Intelligence Examination Without Prepared Blanks. J. Crosby Chapman. *Journal of Educational Research.* December, 1920. 777-786. Description of make-up, scoring, and validity of such tests.

Recent Results Obtained from the Otis Group Intelligence Scale. Stephen S. Colvin. *Journal of Educational Research.* January, 1921. 1-12. Clear and adverse criticism of the norms recommended for the Otis tests. Gives several helpful tables of coefficients correlation between intelligence and their important criteria.

The Measurement of Intelligence. V. A. C. Henmon. *School and Society.* February 5, 1921. 151-157. Vice-presidential address; a survey of recent accomplishment in the intelligence field. Points out strength and weaknesses of our present procedure.

Group Intelligence Tests as a Means of Prognosis in High Schools. I. N. Madsen. *Journal of Educational Research.* January, 1921. 43-52. Typical of articles frequently appearing which give statistical results of using the Army Group Test.

Studies in High-School Procedure—Half-Learning. Henry C. Morrison. *The School Review.* February, 1921. 106-118. General destructive criticism of the use of intelligence tests, and concerning statistical measures such as the probability curve. Educational psychologists and statisticians interested in the use of such measures should read this article.

A Comparative Study of the Intelligence of White and Colored Children. R. A. Schwegler and Edith Winn. *Journal of Educational Research.* December, 1920. 838-848. Summarizes and interprets previous studies; presents good bibliography of 17 items. Conclusions based on Binet tests of 116 Junior High School girls.

A Comparison of Three Methods for Making the Initial Selection of Presumptive Mental Defectives. J. E. Wallace Wallin. School and Society. January 8, 1921. 31-44. Presents valuable correlation data for group and individual tests (Pressey and Binet); argues for the competency of school staffs to estimate the abilities of pupils.

Intelligence and Industrial Tests in Institutional Administration. Edgar A. Doll. Journal of Delinquency. 1920, 5, 215-224.

Practice Effects in Intelligence Tests. Knight Dunlap and Agnes Snyder. Journal Experimental Psychology. 1920, 3, 357-377. Taking the Army Alpha four times increases ability considerably in such works.

Mental Tests. F. N. Freeman. Psychological Bulletin. 1920, 17, 353-363. A critical survey of 48 recent articles.

The Block-Design Tests. S. C. Kohs. Journal Experimental Psychology. 1920, 3, 357-377. Describes a new "performance" test for general intelligence.

Suggestions Looking Toward a Fundamental Revision of Current Statistical Procedure as Applied to Tests. Sydney L. Pressey. Psychological Review. 1920, 27, 466-472.

2. ARTICLES ON THE RATING OF HUMAN CHARACTER.

Character vs. Intelligence in Personality Studies. Guy G. Fernald. Journal Abnormal Psychology. 1920, 15, 1-10. Points out the need of evaluating "personality," "temperament," etc., for supplementing intelligence ratings.

What Should Teacher-Rating Schemes Seek to Measure? Raymond A. Kent. Journal of Educational Research. December, 1920. 802-807. Suggests another scheme for rating teacher as a professional worker.

The Construction of a Teacher-Rating Scale. C. A. Wagner. The Elementary School Journal. January, 1921. 361-366. Concerning new teacher's rating scale and recommends "the use of 'suggestions'" as the unit of measurement of teaching qualifications.

3. MISCELLANEOUS ARTICLES.

Excitability in Delinquent Boys. Mildred S. Covert. Journal of Delinquency. 1920, 5, 224-240.

Personality from the Introspective Viewpoint. Harold I. Gosline. Journal Abnormal Psychology. 1920, 15, 36-45.

The Public Schools and the Treatment of Delinquent Children. Lilburn Merrill. Journal of Delinquency. 1920, 5, 207-215.

Correlation. J. B. Miner. Psychological Bulletin. 1920, 17, 388-396. A summary of 42 articles.

Child Psychology. D. Mitchell. Psychological Bulletin. 1920, 17, 363-775. A summary of 48 recent articles—Educational Psychology. C. T. Gray. Psychological Bulletin. 1920, 17, 375-387. A summary of 103 articles.

The Psychology, Biology and Pedagogy of Genus. L. M. Terman and J. M. Chase. *Psychological Bulletin.* 1920, 17, 397-410. A summary of 95 articles.

Do We Think in Words? Arthur S. Otis. *Psychological Review.* 1920, 27, 399-420. A criticism of the treatment of mental processes by John Watson in 'Psychology from the Standpoint of a Behaviorist.'

The Effect of Fatigue on Attention. John J. B. Morgan. *Journal Experimental Psychology.* 1920, 3, 319-334. Fatigue effects measured by amount learned, retained and recognized as a result of 4 hours' continuous work in memorizing German words.

Effects of Smoking on Mental and Motor Efficiency. Sven Froeberg. *Journal Experimental Psychology.* 1920, 3, 334-347. The moderate use of tobacco appears to have little effect upon mental work.

NEW PUBLICATIONS IN EDUCATIONAL PSYCHOLOGY AND RELATED FIELDS OF EDUCATION

1. *A report on the experimental evaluation of factors which contribute to effective silent reading.* The studies reported in this year-book will no doubt play a significant part in the determination¹ of further research. Several of them are serious attempts at an experimental evaluation of procedure in the teaching of silent reading. W. W. Theisen's contribution is a summary of experimental evidences and current thought concerning fourteen factors affecting results in primary reading. The evidence presented warrants the conclusion that intelligence is probably the most significant factor, although this has never been sufficiently recognized.

Other factors discussed are regular attendance, home influences, amount read, difficulty of material, interest, supervision, and quality of teaching. The studies from which information is drawn are listed in a bibliography. Dean Gray's article deals with the individual difficulties of five pupils in the intermediate grades, and demonstrates the value of careful psychological diagnosis and prescription. John A. O'Brien reports a study of the development of speed in silent reading. Twelve factors are listed. Three types of training were especially devised to incorporate the factors listed. There is a report of a carefully controlled experimental evaluation of the training resulting from its use on approximately 1200 pupils and an interpretation of the results.

The article on motivated drill in silent reading makes practical suggestions, which are much needed in the re-direction of instruction. The value of intense focalization, keen interest and individual activity are demonstrated.

The value of a single reading has, no doubt, been greatly overestimated according to the study reported by A. Yoakum. Suggested methods for careful reading and motivation are submitted in an article by C. E. Germane and Harry A. Greene. Chapters IX

¹The National Society for the Study of Education. Twentieth Yearbook, Part II. *Report of the Society's Committee on Silent Reading.* Bloomington, Ill.: The Public School Publishing Co., 1921. Pp. IX + 172.

and X by J. L. Packer and Daniel Starch respectively, deal with the vocabulary and content of readers.

In Section II of the Yearbook are assembled a number of interesting exercises all of which have seen classroom use. The group of exercises and suggested scale offered by P. R. Heller and S. A. Courtis are an interesting combination of drawing and reading which should appeal to primary children. They suggest the possibilities of improving educational practice by submitting practical suggestions based on psychological research and classroom experimentation.

The chapter by Dr. Burgess is an abstract from a monograph mentioned elsewhere in this issue of the Journal. L. Z.

2. *A comparison of six group intelligence tests.*² This is a very interesting intensive study of six group intelligence scales which were given in the schools of Champaign, Illinois. The chief purpose of the author was to compare and contrast these six scales from the standpoint of their efficiency and adaptability as measuring instruments for general school use. With this end in view, he considers the time consumed in administering the tests, the correlations with scholarship, the correlations between the different scales, the reliability of the total scores and the like. This is all very good and of great interest to the psychologist engaged in mental testing, but much of it would be too technical for the practical worker in the field. It is possible, however, that the practical worker might be led to accept some of the easily comprehended conclusions that the author arrives at rather dogmatically in attempting to summarize his comparisons of the six scales. This summary seems to show the Vocabulary Test (Holley's) superior in more points to the other five scales (Otis, Theisen Classification, Whipple's Group, Pressey Primer, Haggerty Delta 1). The conclusion that the Vocabulary Test is the best test to use in Grades 3 to 12 might easily be made by the uncritical. The critical reader will not be so readily convinced that the author has chosen all the important criteria for judging a scale and he will also note no attempt to evaluate the importance of the criteria chosen. Nevertheless, we need more studies of this type at the present time when group scales of all sorts are appearing

²Holley, Charles E. *Mental Tests for School Use*. Bureau of Educational Research, Bulletin No. 4. University of Illinois, March 8, 1920. Pp. 91.

in great numbers. The emphasis should be laid upon trying to discover for what particular purpose any specific scale is adapted.

In addition to this discussion of the six scales the author presents the results of the survey of the schools of Champaign and his presentation of the data should be useful to superintendents and others who are planning mental surveys of their school system.

The first part of the monograph gives a very readable discussion of the value and uses of mental tests in general, as well as a description of what is meant by general intelligence and what mental tests can and cannot do. This part of the monograph should serve as a splendid introduction to the beginner in the subject.

R. PINTNER.

Ohio State University.

3. *An important treatise on educational measurement.* Dr. May Ayres Burgess's new book, "The Measurement of Silent Reading,"² is of such striking importance that the editors have decided to organize about the discussion of it a critical symposium on educational measurement. This we trust will be published in the May issue. At the present time, therefore, we supply merely a brief statement of what Dr. Burgess has done. The book makes three types of contribution to education. First, it reports a new scale for measuring ability in silent reading. Next, it presents a careful critique of the construction of scales in silent reading. Third, it is a very careful analytical study of the construction of educational scales in general. Fundamental issues of technical scale construction and of educational thinking have been raised—issues which should be debated vigorously.

H. O. R.

4. *Presentation of a history test.* A recent monograph by Dr. M. J. Van Wagenen⁴ points the way to much more objective methods of testing in the field of United States history. This study describes how questions involving facts, thought and character judgment were standardized so that these commonly emphasized aims could be accurately measured.

²Burgess, M. A. *The Measurement of Silent Reading*. New York: Russell Sage Foundation, 1921. Pp. 163.

⁴Van Wagenen, M. J. *Historical Information and Judgment in Pupils of the Elementary School*. Teachers College, Columbia University, Contributions to Education, No. 101.

These questions were derived from a preliminary test given to 1200 children in three public schools of New York City in grades 4-8 and from revised tests taken by 2000 more pupils.

Van Wegenen lists five desiderata for measuring historical abilities: First, that "the tests be symptomatic of important abilities really desired by the schools." The outcomes tested are important ones, though facts are valuable mainly for use in solving problems, in making judgments and in interpreting historical situations. It is to be hoped that this investigation will stimulate others to complete the standardization of the objectives of history teaching. Only in this way can we obtain a scientific statement of what the aims of historical instruction are. His second thesis is that "the tests be not too much disturbed by linguistic difficulties so that ability in history, not in reading and composition, may be chiefly measured." His results, in general, accord with this desideratum. Third, "that the measurement of a small group, such as a class of twenty-five or more, may be made with sufficient precision." A correlation of above .7 between each type of test is obtained. Also detailed tables of the degree of difficulty of the questions with a "key" for the scoring of the same makes possible accurate measurement of a class. However, if he would also transmute his scores into the familiar percentage scheme with the value of each question on that basis indicated on the teacher's scorecard, it would be much more helpful to the history teacher. His fourth point that "the tests be capable of extension to alternative forms so as to reduce harm done by coaching" is taken care of in these scales by including 70 Fact, 44 Thought and 40 Judgment questions. These together well exhaust the content of United States History. Fifth, the last criterion that "the administration and scoring of the tests be convenient" is illustrated by the fact that fifty minutes is sufficient to exhaust the abilities of all save a few and that the scoring can be done relatively accurately by the average history teacher by means of his "key." E. U. RUGG.

Teachers College, Columbia University.

5. *A report of use of tests.*⁵ The students of a summer course in experimental education under the direction of Baldwin were each given a specific piece of work in tests and measurements. These

⁵Baldwin, Bird T., and others. *Studies in Experimental Education*. The Johns Hopkins University Studies in Education, No. 3. Baltimore, 1920. Pp. 75.

experiments were conducted by the students in a small demonstration school of six grades containing 129 pupils "who represent, in most instances, examples of maladjustment in educational progress." The tests conducted by these university students were written up and form eleven articles in the monograph. These studies give the results of the following tests: Yerkes Scale, Terman Scale, Courtis Arithmetic, Woody Arithmetic, several handwriting scales, Kansas and Starch Reading, several spelling scales, Trabue Completion, Hillegas and Ballou Composition.

Naturally the results obtained on a small group of children under the circumstances described could not be expected to be of great importance. No particularly new line of attack is followed in any of the studies. Some of them stress the data obtained and others attempt some discussion or criticism of the tests themselves. Some of the studies hardly go further than a clear presentation of the data. It might rightly be questioned whether students' themes of this type warrant publication in a university monograph or, indeed, whether they warrant publication at all. To the reviewer what justification there is would seem to lie mainly in the stimulus to advanced study and independent research which such an undertaking might give the students concerned. There can be no doubt that a group of students working towards the publication of their studies would be stimulated thereby. The whole monograph, therefore, forms a splendid example of the project method applied to the study of mental and educational tests. It shows moreover how well adapted to the project method is the study of this phase of educational psychology. In this field individual and group projects can be used very effectively.

The first chapter of the monograph is a summary of the students' work by Baldwin himself, in which he attempts to correlate all the separate studies. The presentation of a great many correlations of such a heterogeneous and small group of pupils does not help us very much. The developmental graph of the circular type combining all the various ratings of a particular child is interesting.

The monograph should be read by all teachers of university classes in tests and measurements in order to get some idea of how the project method may be applied to their classes. R. PINTNER.

Ohio State University.

6. *Psychological factors involved in social reconstruction.* That the facts and principles of "dynamic" psychology are being studied and applied by students of sociology is clearly evident in a recent book by Professor Patrick.⁶ Following somewhat the trend of the late Carleton Parker, the author amplifies the thesis that society must be reorganized "not on a basis of adequate scale of living, material comforts, wealth or efficiency" but "to satisfy instinctive needs" and to "furnish a field for human activities." Accepting, in the main, the fundamental impulses or "instincts" listed by James, McDougall and Thorndike, the author considers, in terms of these, the motives behind many present day activities ranging from the movies, dancing and smoking to crime, war and anarchy. The human animal, driven by numerous inborn impulses, fails to satisfactorily exercise them under conditions of modern industrial and social conditions, consequently a myriad of indirect expressions,—“sublimations”—of the undesirable sort appear. In general, it will be gratifying to many, that Professor Patrick has handled his thesis without resort to the fantastic explanations of the Freud-Jung-Adlertype. He restricts his principles rather well to the recognized findings of science, carrying the interpretations to remote fields, to be sure. The psychologist will object to a neglect of established facts of individual differences. Professor Patrick's humans as assembled, are too much of a mould; in particular they are too similar to the exceptional man in their interest in thinking, problem solving and creative work. In social reorganization, what is urged is the adaptation of the environment and work to the needs of the man, and what is needed is a scientific study of the "wants" of man to discover which may be changed through education, which persist in whatever environment. The book is important because it rebels against the "strange forgetfulness of the fact that however important social and political readjustments may be, the world cannot be made over as long as human material—the minds and bodies of men—remain the same."

A. I. G.

7. *A Yearbook of New Project Materials.* Part I of the Twentieth Yearbook of the National Society for the Study of Education⁷

⁶Patrick, George Thomas White. *The Psychology of Social Reconstruction* Boston: Houghton Mifflin Company, 1920. Pp. IX + 273.

⁷The Twentieth Yearbook of the National Society for the Study of Education, Part I. *Second Report of the Society's Committee on New Materials of Instruction.* Bloomington, Ill.: Public School Publishing Co., 1921. Pp. XV + 237.

consists of a collection of 285 projects compiled by a committee of eleven who drew upon more than 100 members of subcommittees. The materials, classified by grades from the kindergarten to the sixth grade, inclusive, are given 115 pages, in which 193 projects are described. Projects numbering eighty are classified by subjects for the Junior High School. Ten projects for special classes are listed. The book includes an annotated bibliography of 404 titles, classified under 38 headings. It is by far the best collection of concrete samples of project teaching which has appeared. A. I. G.

8. *The Project Method in Education*. A Riverside Educational Monograph,⁸ deals in an intelligible manner with the theory of teaching by projects, with a series of illustrative "projects" in history, civics, manual arts, agriculture, current events, etc. A project "is no more or no less than the natural concrete expression of modern principles of education in practice." What the modern principles of education are is shown in the first three chapters by tracing their development from Rousseau through Pestalozzi, Froebel, Herbert, and in turn the growth of their influence in America, eventuation in a brief exposition of the doctrines of James, Thorndike, Dewey, McMurry and others. "Development through natural processes," "natural interaction with the environment," "self expression through activity," "learning to do by doing," "self activity at work solving problems," etc., are general terms rather frequently repeated which the author attempts to illustrate in some detail. Compared to certain recent writings, Mr. Stockton's recommendations are conservative. He makes no fiery denunciation of existing conditions. He does not recommend the abolition of established school subjects although he would seek more fruitful attack upon them. "It is safe to say that the keynote of the new spelling [for example] is that spelling must be definitely *taught*—that mere 'exposure' is too uncertain." He would merge several subjects; e. g., "geography, elementary science, nature-study, school-gardening, history, civics, current events, and primitive life may be reduced without loss to history and geography." He believes in adult interference in a case where we cannot expect that the child "will absolutely form his own

⁸Stockton, James Leroy. *Project Work in Education*. Boston: Houghton Mifflin Company, 1920. Pp. XIV + 167.

opinion about it," and if he persist in undesirable conduct "he is finally even put into restraint."

Perhaps most important is the recommendation that we must begin "really to teach children how to study, through making children themselves conscious of the definite factors involved." "It is no longer possible to teach all of the multitudes of facts accumulated by the race; but it is possible to develop 'methods of attack' which will make the student independent in his solutions of difficulties as they arrive. A good course in 'How to Study' or 'How to Think' fulfills this purpose." If Dr. Stockton is a little optimistic in his claims for the methods he describes, he nevertheless succeeds in stating rather effectively a good many issues. The book will render service as a very summary introduction to modern educational theory and practice; it is too summary, of course, to be useful as a concrete guide to practice.

A. I. G.

9. *A New Spelling Book.* Research upon the problems of spelling is perhaps producing a closer approximation to a scientific subject-matter and method than in any other fundamental study. For a number of years Professors Horn and Ashbaugh have been conducting investigations in both fields and their spelling book,⁹ just off the press, is unquestionably more scientifically constructed than any which has yet appeared. The vocabulary consists of a minimum list of 3998 words which appeared most frequently in some 700,000 running words of correspondence, the result of nine different studies. Supplementary lessons for each grade include 580 additional words. The words are apportioned to each of the eight grades, with the difficulty (standard number of errors for each column of 10 or 20) given at the foot of the column. The grading of words is based upon three factors; 1) the frequency of use in correspondence; 2) the difficulty of the words as shown on the Ashbaugh Scale, and 3) the frequency with which the word appears in representative first, second and third grade readers. An important addition to the usual spelling scale is a list of abbreviations and names of states, common measures and conventional commercial and professional titles. A list of 50 "demons"—the most difficult words of the Ashbaugh scale—are found in a series of dictation letters at the end of the book.

⁹Horn, Ernest, and Ashbaugh, Ernest J. *Lippincott's Horn-Ashbaugh Spelling Book*. Philadelphia: I. B. Lippincott Company, 1920. Pp. XX + 105.

While the grading, especially above grade IV, is not perfect, the list as a whole is a distinct improvement over others now available.

The method of teaching and learning to spell, essentially as reported by Professor Horn in the 18th Yearbook of the National Society for the Study of Education, is distinct departure from convention. The words are listed in columns of 20 each (10 in grade I). There are no jingles, no pictures, no groupings, no rules, no underlining of difficult parts, no heavy face type, no stunts of any sort. One method is described for learning to spell a word. The pupil masters the method and then attacks a word which is learned through sheer habituation. It is assumed that the learning of each word is specific. Each child is enabled by a weekly system of tests and reviews to attack only those words which offer difficulty. Provision is made for systematic review on Monday, Wednesday and Friday, two reviews one month later and at the beginning of each year the 80 most difficult words of the preceding grade are reviewed. Other details cannot be given. The method in a general way is an illustration of certain present-day principles of psychology which considers learning to be rather specific in character. A. I. G.

10. *The development of a new economic and social curriculum in our secondary schools.* It is difficult—yes, almost impossible—for a teacher of social science to find adequate printed material to put in the hands of high school pupils. For years, we have recognized that new material would be introduced into our “social-studies” curriculum only through new types of text books. New books come from the press constantly. In the main they are formal and contain “structural” (as opposed to functional) discussions of government, administrative history, and formal economics. Two books have appeared recently,¹⁰ two published within the month which typify the different movements in this field.

Our Economic Organization by Professors Marshall and Lyon may well be regarded as an epoch-marking book. It represents a completely new type of economics and attempts to do for the high

¹⁰Marshall, L. C., and Lyon, L. S. *Our Economic Organization*. New York: MacMillan Company, 1921. Pp. X + 503.
Ames, E. W., and Eldred, A. *Community Civics*. New York: MacMillan Company, 1921. Pp. XIV + 387.

school what is being done in various books of "readings" of industrial society by such men as Professors Marshall, Hamilton, and the like. This new book aims to give pupils a clear notion of the working of economic laws and of the building and carrying on of industrial society, by means of concrete descriptions and interpretations of the way in which we live and work in an industrial society. It is a radical departure from the common run of "economics" textbooks, in its emphasis upon the use and work of industry—banks, business organizations, government, scientific management, and the like. The authors have broken away from tradition in many places; for example, in the complete omission of discussions of "value" and "distribution."

Not only is the book an improvement from the standpoint of the selection of material, but also from the psychological viewpoint of arrangement. Historical backgrounds are given by a series of fairly detailed and sharp contrasts. This is one of the exemplifications of the theory of presenting historical development which the present reviewer is developing and hopes to see carried out in the organization of our historical courses. Furthermore, the writers have broken away from the encyclopedic, paragraphic treatment to which nearly all school text books are committed. At the beginning of each "study" ("studies" replace "chapters") the exact purpose of the work is stated definitely. The authors likewise present long lists of problems, questions, activities, for teachers to use with their pupils. We believe that learning could be more enhanced if the material of the book had been *organized around* problems and questions. We doubt if public school teachers are equipped to make the most effective use of lists of problems, questions, and activities, which are appended to chapter discussions.

The newest addition to the list of community civics books is one by Ames and Eldred. This is a curious mixture of concreteness in presentation of certain problems such as, the beginning of a community; home, family, and the community; and a formal, abstract, definitional treatment of many other community activities. For example, in the chapter on migration, the detailed, and interesting "story of Pietro" is to be contrasted with the brief paragraphs on the "causes" and "results of immigration." The book is a collection of textlike chapters on different community activities. The present reviewer is utterly unable to find any continuity of organization or

sequential thread running through the presentation. Several chapters appear on education, public health, protection of life and property, work and play, correction, migration, the needy and dependent, government and making a living, the lawmakers of state and nation and the like.

H. O. R.

11. *A concrete methods book for Junior-High School English.* The supervised study movement is being kept alive by such devices as new textbooks in the supervision of study in the various school subjects. MacMillan is issuing a series of books under such captions as: *Supervised Study*, Miss Simpson's *Supervised Study in History*, and *Supervised Study in English*, all under the editorship of Professor A. L. Hall-Quest. His book, *Supervised Study*, was published several years ago. A number of public school teachers, notably in Rochester, New York, have been preparing books of technic on the teaching of the separate subjects, following closely the principles laid down by the editor.

This is well illustrated by Miss McGregor's¹¹ application in the field of English. The book gives lesson plans. It is full of suggestions for teaching technic from the experience of a good teacher; it suggests new materials, and presents digests of helpful devices for teaching the subject; *e. g.* "corrective substitutes," discussions of the correct use of verb forms, and the like. In each case the writer tells what was done by the teacher under her supervision. Thus, the book is an empirical description of what has apparently worked in one school system. As a book of suggestions on methodology, it should be helpful.

The present reviewer believes that it is possible to improve the teaching of the school subjects more directly than by recommending to teachers that the hours of a school program be divided up into different periods in which different tasks are assigned to different periods. This is a recommendation of these supervised study workers. For example, the Supervised Study advocates recommend (as is illustrated by Professor Hall-Quest and his co-workers) that each class exercise should be sub-divided and definite provision made for different kinds of work. For example for a lesson in oral com-

¹¹McGregor, L. A. *Supervised Study in English*. New York: MacMillan Company, 1921. Pp. XII + 220.

position in the 7th grade, covering 50 minutes, it is recommended that "review" receive 9 minutes, "assignment" 10 minutes, "silent study" 6 minutes, "oral expression" 20 minutes, and "summary" 5 minutes. The application to literature, to written composition, and to grammar is made with distinctly different recommendations as to division of time. An experiment which the reviewer is now carrying on in fifth grade social studies raises questions about the validity of this procedure and suggests that to assign periods of different kinds of work within the class hour may quite completely formalize the lesson.

H. O. R.

12. *A Significant Conference Report on Sex Education.* The booklet¹² referred to is a preliminary edition of the findings of a series of conferences on sex education held under the auspices of the International Committee of Young Men's Christian Association. The work was done under the active direction of Dr. Thomas W. Galloway. The method of the conference is indicative of the trustworthiness of a report in which dogmatism, prejudice, and unsupported personal opinion give way to the unbiased group judgment of a large number of conferees whose special interests and abilities fit them to contribute from widely separated viewpoints to a subject of general concern. The reader is impressed by the potential significance of the co-operation of general educators, teachers, biologists, psychologists, psycho-analysts, sociologists, physicians, and physical directors, and feels it incumbent upon himself to assume a similarly impersonal attitude in evaluating the report.

Because so many people still question whether anything can or ought to be attempted in sex education, and others over-emphasize partial and emergency objectives, ignoring other equally important aims, the issue is carefully defined. "It is quite obvious that sex qualities and relations profoundly influence conduct, development, and relations which are not sexual; and that non-sexual phenomena equally influence the sexual attitudes and conduct. Hence this statement of imbedding sex education and its motives and purposes in the whole purpose of education and life was the most conspicuous point of agreement in the conference, in its effort to outline objectives. Sex education is not in its nature a special form

¹²The American Social Hygiene Association. *Preliminary Synthesis and Integration of the Returns of the Sex Education Conference.* New York, 1920. Pp. 95.

of education; it is merely character education which ceases to ignore one of the most imperious and pervasive groups of incentives in human nature."

The proposed integration takes account of individual differences but recognizes the significance of a psychological gradation of the proposed training and motivation.

The special report of the General Education Section of the conference shows general agreement on the following point among many others: "Anything which segregates sex education arouses undue consciousness and more or less apprehension, which interferes with the best emotional results."

The report deals rather exclusively with the problems relating to boys and young men. Revision and restatement with reference to the needs of girls and young women is contemplated. L. Z.

III. ADDITIONAL PUBLICATIONS RECEIVED.*

A. MENTAL AND EDUCATIONAL TESTS.

ROBACK, A. A. *Roback Mentality Tests for Superior Adults*. Boston: A. A. Roback, Simmons College. 1920.

B. PUBLICATIONS IN THE GENERAL EDUCATIONAL FIELD.

COURSALT, J. H. *The Principles of Education*. Boston: Silver, Burdett & Co. 1920. Pp. XII + 468.

HOSIC, JAMES F. *Sample Projects*. First Series. 506 W. 59th street, Chicago. 1920. Pp. 32.

KELLY, R. W. *Training Industrial Workers*. New York: Ronald Press Co. 1920. Pp. XXI + 437.

PATRICK, G. T. W. *The Psychology of Social Reconstruction*. Boston: Houghton-Mifflin Company. 1920. Pp. IX + 273. \$2.

SNEDDEN, D. *Sociological Determination of Objectives in Education*. Philadelphia: J. B. Lippincott & Co. 1921. Pp. 322. \$2.50.

STRAYER AND ENGELHARDT. *The Classroom Teacher*. New York: American Book Co. 1920. Pp. 400.

TURNER, E. A. *The Essentials of Good Teaching*. New York: D. C. Heath & Co. 1920. Pp. XIII + 271.

C. NEW SCHOOL TEXTBOOKS.

BEARD AND BAGLEY. *A First Book in American History*. New York: Macmillan Company. 1920. Pp. IX + 460.

LINDQUIST, T. *Junior High School Mathematics*. Book I. New York: Scribner. 1920. Pp. VI + 235. \$1.12. Book II. Pp. VII + 237. \$1.24. Book III. Pp. VII + 243. \$1.36.

D. PUBLICATIONS OF UNITED STATES BUREAU OF EDUCATION.

A Survey of Education in Hawaii. Bureau of Education. Bulletin No. 16. 1920. Pp. V + 408.

E. MISCELLANEOUS PUBLICATIONS.

California: *Report of the Special Legislative Committee on Education*. Sacramento: California State Printing Office. 1920. Pp. 96.

COFFMAN, L. W. *Teacher Training Departments in Minnesota High Schools*. New York: General Education Board. 1920. Pp. VII + 92.

General Education Board: *Annual Report 1919-1920*. New York. Pp. VII + 141.

GOODSPEED, HELEN C. *Suggestions for Teaching Homemaking in the Grades and High School*. Wisconsin: C. P. Gary, State Superintendent. 1920. Pp. 39.

HALL, A. B. *Dynamic Americanism*. Indianapolis: The Bobbs-Merrill Company. 1920. Pp. 325.

RICH, FRANK M. *School Economies*. Baltimore: Warwick & York. 1920. Pp. 72.

*Publications which are reviewed in this issue are not listed here.

THE JOURNAL OF EDUCATIONAL PSYCHOLOGY

Vol. XII

APRIL, 1921

No. 4

THE PSYCHOLOGY OF DRILL IN ARITHMETIC: THE AMOUNT OF PRACTICE*

E. L. THORNDIKE.

Teachers College, Columbia University.

THE AMOUNT OF PRACTICE.

It will be instructive if the reader will perform the following experiment as an introduction to the discussion of this article before reading any of the discussion.

Suppose that a pupil does all the work, oral and written, computation and problem solving, presented for grades I to VI inclusive (that is, in the first two books of a three-book series) in the average text-book now used in the elementary school. How many times will he have exercised each of the various bonds involved in the four operations with integers. That is, how many times will he have thought "1 and 1 are 2," "1 and 2 are 3," etc., etc. Every case of the action of each bond is to be counted.

Since estimating for the entire series is too long a task, it will be sufficient to use eight or ten from each, say:

3 + 2	13, 23, etc., + 2	7 + 2	17, 27, etc., + 2
" 3	" 3	" 3	" 3
" 4	" 4	" 4	" 4
" 5	" 5	" 5	" 5
" 6	" 6	" 6	" 6
" 7	" 7	" 7	" 7
" 8	" 8	" 8	" 8
" 9	" 9	" 9	" 9

*Copyright, 1921, by E. L. Thorndike.

3—3	7—7	9×7	$63 \div 9$
4 “	8 “	7×9	64 “
5 “	9 “		65 “
6 “	10 “		66 “
7 “	11 “	8×6	67 “
8 “	12 “	6×8	68 “
9 “	13 “		69 “
10 “	14 “		70 “
11 “	15 “		71 “
12 “	16 “		

TABLE I.

Estimates of the amount of practice provided in Books I and II of the average three-book text in Arithmetic; by 50 experienced teachers.

Arithmetical fact.		Estimate.			Range required to include half of the estimates.
		Lowest.	Median.	Highest.	
3 or 13 or 23, etc., + 2		25	1,500	1,000,000	800—5,000
“	3	24	1,450	80,000	475—5,000
“	4	23	1,150	50,000	750—5,000
“	5	22	1,400	44,000	700—5,000
“	6	21	1,350	41,000	700—4,500
“	7	21	1,500	37,000	600—4,000
“	8	20	1,400	33,000	550—4,100
“	9	20	1,150	28,000	650—4,500
7 or 17 or 27, etc., + 2		20	1,250	2,000,000	600—5,000
“	3	19	1,100	1,000,000	650—4,900
“	4	18	1,000	80,000	650—4,900
“	5	17	1,300	80,000	650—4,400
“	6	16	1,100	29,000	650—4,500
“	7	15	1,100	25,000	500—4,500
“	8	13	1,100	21,000	650—3,800
“	9	10	1,275	17,000	500—4,000
3—3		25	1,000	100,000	500—4,000
4 “		20	1,050	500,000	525—3,000
5 “		20	1,100	2,500,000	650—4,200
6 “		10	1,050	21,000	650—3,250
7 “		22	1,100	15,000	550—3,050
8 “		21	1,075	15,000	650—3,000
9 “		21	1,000	15,000	700—2,600
10 “		20	1,000	20,000	600—2,500
11 “		20	1,000	15,000	465—2,550
12 “		18	1,000	15,000	650—2,100
7—7		10	1,000	18,000	425—3,000
8 “		15	1,000	18,000	413—3,100
9 “		15	950	18,000	550—3,000
10 “		15	950	18,000	600—3,950
11 “		10	900	18,000	550—3,000
12 “		10	925	18,000	525—3,100
13 “		10	900	18,000	500—2,600
14 “		10	900	18,000	500—3,100
15 “		10	925	18,000	500—3,000
16 “		10	875	18,000	500—2,500

TABLE I—(Continued.)

Arithmetical fact.	Estimate.			Range required to include half of the estimates.
	Lowest.	Median.	Highest.	
9×7	10	700	20,000	500—2,000
7×9	10	700	20,000	500—1,750
8×6	10	750	20,000	500—2,500
6×8	9	700	20,000	500—2,500
$63 \div 9$	9	500	4,500	300—2,500
64 “	9	200	4,000	100— 700
65 “	8	200	4,000	100— 600
66 “	7	200	4,000	100— 550
67 “	7	200	4,000	75— 450
68 “	6	200	4,000	87— 575
69 “	6	200	4,000	87— 450
70 “	5	200	4,000	75— 575
71 “	5	200	4,000	75— 700

Having made his estimates the reader should compare them first with similar estimates made by experienced teachers (shown in Table I), and then with the results of actual counts for representative text-books in arithmetic (shown in Tables II to VII).

It will be observed in Table I that even experienced teachers vary enormously in their estimates of the amount of practice given by an average text-book in arithmetic, and that most of them are in serious error by over-estimating the amount of practice. In general it is the fact that we use text-books in arithmetic with very vague and erroneous ideas of what is in them, and think they give much more practice than they do give.

The authors of the text-books as a rule also probably had only very vague and erroneous ideas of what was in them. If they had known, they would almost certainly have revised their books. Surely no author would intentionally provide nearly four times as much practice on $2 + 2$ as on $8 + 8$, or eight times as much practice on 2×2 as on 9×8 , or eleven times as much practice on $2 - 2$ as on $17 - 8$, or over forty times as much practice on $2 \div 2$ as on $75 \div 8$ and $75 \div 9$, both together. Surely no author would have provided intentionally only twenty to thirty occurrences each of $16 - 7$, $16 - 8$, $16 - 9$, $17 - 8$, $17 - 9$, and $18 - 9$ for the entire course through grade VI; or have left the practice on $60 \div 7$, $60 \div 8$, $60 \div 9$, $61 \div 7$, $61 \div 8$, $61 \div 9$, and the like to occur only about once a year!

Tables II to VII show that even gifted authors make instruments for instruction in arithmetic which contain much less practice on certain elementary facts than teachers suppose; and which contain relatively much more practice on the more easily learned facts than on those which are harder to learn.

How much practice should be given in arithmetic? How should it be divided amongst the different bonds to be formed? Below a certain amount there is waste because the pupil will need more time to detect and correct his errors than would have been required to give him mastery. Above a certain amount there is waste because of unproductive over-learning. If 668 is just enough for 2×2 , 82 is not enough for 9×8 . If 82 is just enough for 9×8 , 668 is too much for 2×2 .

TABLE II.

Amount of Practice: Addition Bonds in a Recent Text-book (A) of Excellent Reputebooks I and II. All save certain supplementary material which does not appear to be considered a part of the normal work.

The Table reads: " $2 + 2$ was used 226 times, $12 + 2$ was used 74 times, $22 + 2$, $32 + 2$, $42 + 2$ and so on were used 50 times."

	2	3	4	5	6	7	8	9	Total.
2	226	154	162	150	97	87	66	45	
12	74	53	76	46	51	37	36	33	
22, etc.	50	60	68	63	42	50	38	26	
3	216	141	127	89	82	54	58	40	
13	43	43	60	70	52	30	22	18	
23, etc.	15	30	51	50	42	32	29	30	
7	85	90	103	103	84	81	61	47	
17	35	25	42	32	35	21	29	16	
27, etc.	30	23	32	29	24	23	25	28	
8	185	112	146	99	75	71	73	61	
18	28	35	52	46	28	29	24	14	
28, etc.	53	36	34	38	23	36	27	27	
9	104	81	112	96	63	74	58	57	
19	13	11	31	38	25	14	22	11	
29, etc.	19	17	27	20	32	32	19	18	
2, 12, 22, etc., +	350	277	306	260	190	174	140	104	1,801
3, 13, 23, etc.	274	214	230	209	176	116	109	88	1,406
7, 17, 27, etc.	148	138	187	164	141	125	115	91	1,109
8, 18, 28, etc.	266	183	232	185	126	136	124	102	1,354
9, 19, 29, etc.	136	109	170	154	120	120	99	86	994
Totals..	1,164	921	1,125	972	753	671	687	471	

It is possible to find the answers to these questions for the pupil of median ability (or any stated ability) by suitable experiments. The amount of practice will of course vary according to the ability of the pupil. It will also vary according to the interest aroused in him and the satisfaction he feels in progress and mastery. It will also vary according to the amount of practice of other related bonds; $7 + 7 = 14$ and $60 \div 7 = 8$ and 4 remainder will help the formation of $7 + 8 = 15$ and $61 \div 7 = 8$ and 5 remainder. It will also of course vary with the general difficulty of the bond, $17 - 8 = 9$ being under ordinary conditions of teaching, harder to form than $7 - 2 = 5$.

Until suitable experiments are at hand we may hazard estimates for the fundamental bonds as follows, assuming that by the end of grade VI a strength of 199 correct out of 200 is to be had, and that the teaching is by an intelligent person working in accord with psychological principles as to both ability and interest.

For one of the easier bonds, most facilitated by other bonds (such as $2 \times 5 = 10$, or $10 - 2 = 8$, or the double bond $7 =$ two threes and 1 remainder) in the case of the median or average pupil, twelve practices in the week of first learning, supported by twenty-five practices during the two months following, and maintained by thirty practices well spread over the later periods. For the more gifted pupils, lesser amounts down to six, twelve and fifteen may suffice. For the less gifted pupils, more may be required up to thirty, fifty and a hundred. It is to be doubted, however, whether pupils requiring so much sheer repetition of these easy bonds as this should be taught arithmetic beyond a few matters of practical necessity.

For bonds of ordinary difficulty, with average facilitation from other bonds (such as $11 - 3$, 4×7 , or $48 \div 8 = 6$) in the case of the median or average pupil, twenty practices in the week of first learning, supported by thirty during the two months following, and maintained by fifty practices well spread over the later periods. Gifted pupils may gain and keep mastery with twelve, fifteen and twenty practices respectively. Pupils dull at arithmetic may need up to twenty, sixty and two hundred. Here, again, it is to be doubted whether a pupil for whom arithmetical facts, well-taught and made interesting, are so hard to acquire as this, should learn many of them.

TABLE III.

Amount of Practice: Subtraction Bonds in a Recent Text-book (A) of Excellent Repute: Books I and II. All save certain supplementary material, which does not appear to be considered a part of the normal work.

Frequencies of subtractions of 1 from 1, 1 or 2 from 2, 2 or 3 from 3, etc.

		Subtrahends.								
		1	2	3	4	5	6	7	8	9
Minuends	1.....	372								
	2.....	214	311							
	3.....	136	149	189						
	4.....	146	142	103	205					
	5.....	171	91	92	164	136				
	6.....	80	59	69	71	81	192			
	7.....	106	57	55	67	59	156	80		
	8.....	73	50	50	75	50	62	48	152	
	9.....	71	75	54	74	48	55	55	124	133
	10.....	261	84	63	100	193	83	57	124	91
	11.....		48	31	50	36	41	32	46	35
	12.....			48	77	57	51	35	80	30
	13.....				35	22	40	29	35	28
	14.....					25	37	36	49	32
	15.....						33	19	48	20
	16.....							16	36	26
	17.....								27	20
	18.....									19
Totals, Excluding 1 — 1, 2 — 2, etc.		1,258	755	565	613	571	558	327	569	301

TABLE IV.

Frequencies of subtractions not included in Table III.

These are cases where the pupil would, by reason of his stage of advancement, probably operate 35—30, 46—46, etc., each as one bond.

	Subtrahends.									
	1	2	3	4	5	6	7	8	9	
	11	12	13	14	15	16	17	18	19	10
	21	22	23	24	25	26	27	28	29	20
	etc.	etc.	etc.	etc.	etc.	etc.	etc.	etc.	etc.	etc.
10, 20, 30, 40, etc.....	11	29	16	52	32	51	7	30	22	60
11, 21, 31, 41, etc.....	42	14	22	32	12	26	19	52	17	10
12, 22, 32, 42, etc.....	47	97	5	13	9	21	11	24	19	17
13, 23, 33, 43, etc.....	7	40	7	14	15	13	19	19	22	3
14, 24, 34, 44, etc.....	8	28	14	58	13	16	14	26	19	7
15, 25, 35, 45, etc.....	21	28	29	54	51	15	21	12	24	8
16, 26, 36, 46, etc.....	5	18	12	27	35	69	13	17	19	2
17, 27, 37, 47, etc.....	5	9	12	40	32	54	24	12	12	1
18, 28, 38, 48, etc.....	2	16	10	23	22	36	18	47	16	0
19, 29, 39, etc.....	5	7	7	10	13	28	14	23	16	0
Totals.....	153	276	134	323	234	329	160	261	186	117

TABLE V.

Amount of Practice: Multiplication Bonds in Another Recent Text-book (B) of Excellent Repute. Books I and II.

	Multiplicands.									
	0	1	2	3	4	5	6	7	8	9 Totals.
1.....	299	534	472	271	310	293	261	178	195	99 2,912
2.....	350	644	668	480	458	377	332	238	239	155 3,941
3.....	280	487	509	388	318	302	247	199	227	152 3,109
4.....	186	375	398	242	203	265	197	163	159	93 2,281
5.....	268	359	393	234	263	243	217	192	197	114 2,480
6.....	180	284	265	199	196	191	148	169	165	106 1,923
7.....	135	283	277	176	187	158	155	121	145	118 1,755
8.....	137	272	292	175	192	164	158	157	126	126 1,799
9.....	71	173	140	122	97	102	101	100	82	110 1,098
Totals	1,906	3,411	3,414	2,287	2,224	2,095	1,836	1,517	1,535	1,073

TABLE VI.

Amount of Practice: Divisions without Remainder in Text-book (B), Parts I and II.

Integral Multiples of 2 to 9 in sequence, i. e., $4 \div 2$ occurred 397 times; $6 \div 2$ occurred 256 times; $6 \div 3$, 224 times; $9 \div 3$, 124 times.

Divisors.								
2	3	4	5	6	7	8	9	Totals.
397	224	250	130	93	44	98	23	1,259
256	124	152	79	28	43	61	25	768
318	123	130	65	50	19	39	19	763
258	98	86	105	25	24	34	20	650
198	49	76	27	22	30	33	16	451
77	54	36	31	28	27	16	9	278
180	91	50	38	17	13	22	16	427
69	46	37	24	12	17	16	15	236
Totals.....	1,753	809	817	499	275	217	319	142

For bonds of greater difficulty, less facilitated by other bonds (such as $17 - 9$, 8×7 , or $12\frac{1}{2}\%$ of $= \frac{1}{8}$ of), the practice may be from ten to a hundred percent more than the above.

UNDER-LEARNING AND OVER-LEARNING.

Accepting the above provisional estimates as reasonable, we may consider the harm done by giving less and by giving more than these reasonable amounts. Giving less is indefensible. The pupil's time is wasted in excessive checking to find his errors. He is in danger of being practiced in error. His attention is diverted from the learning of new facts and processes by the necessity of thinking out these supposedly mastered facts. All new bonds are harder to learn than they should be because the bonds which should facilitate them are not strong enough to do so. Giving more does harm to

TABLE VII—(Continued.)

Dividend	35	36	37	38
Divisor	4 5 6 7 8 9	4 5 6 7 8 9	4 5 6 7 8 9	4 5 6 7 8 9
No. of occurrences.....	10 31 5 24 5 3	37 16 22 2 6 19	12 8 7 5 3 9	7 8 7 1 1 5
Dividend	39	40	41	42
Divisor	4 5 6 7 8 9	5 6 7 8 9	5 6 7 8 9	5 6 7 8 9
No. of occurrences.....	4 3 7 4 3 1	38 9 2 34 2	6 6 3 7 5	7 28 30 10 3
Dividend	43	44	45	46
Divisor	5 6 7 8 9	5 6 7 8 9	5 6 7 8 9	5 6 7 8 9
No. of occurrences.....	7 5 10 3 3	7 6 4 5 0	24 6 7 10 20	3 3 2 2 2
Dividend	47	48	49	50
Divisor	5 6 7 8 9	5 6 7 8 9	5 6 7 8 9	5 6 7 8 9
No. of occurrences.....	6 2 2 0 3	7 17 4 33 2	4 7 27 9 2	4 6 3 8 2 3 1 2
Dividend	52	53	54	55
Divisor	6 7 8 9	6 7 8 9	6 7 8 9	6 7 8 9
No. of occurrences.....	5 5 5 3	4 3 2 2	12 5 1 16	5 3 4 2 0 13 16 8
Dividend	57	58	59	60
Divisor	6 7 8 9	6 7 8 9	6 7 8 9	6 7 8 9
No. of occurrences.....	0 3 1 3	2 2 3 1	2 3 0 3	3 9 1 1 2 5
Dividend	62	63	64	65
Divisor	7 8 9	7 8 9	7 8 9	7 8 9
No. of occurrences.....	4 6 1	17 5 9	5 22 0	1 10 1 2 1 4 0 1 1
Dividend	68	69	70	71
Divisor	7 8 9	7 8 9	7 8 9	7 8 9
No. of occurrences.....	1 3 2	0 6 1	6 2 1 0	16 10 7 5 3 3 5 3
Dividend	76	77	78	79
Divisor	8 9	8 9	8 9	8 9
No. of occurrences.....	3 2	3 0	4 1 0 2	4 15 2 4 1 2
Dividend	86	87	88	89
Divisor	9	9	9	9
No. of occurrences.....	0	3	2	7

some extent by using up time that could be spent better for other purposes, and (though not necessarily) by detracting from the pupil's interest in arithmetic. In certain cases, however, such excess practice and over-learning are actually desirable. Three cases are of special importance.

The first is the case of a bond operating under a changed mental set or adjustment. A pupil may know 7×8 adequately as a thing

285

by itself, but need more practice to operate it in 7 where he has

to remember that 3 is to be added to the 56 when he obtains it, and that only the 9 is to be written down, the 5 to be held in mind for later use. The practice required to operate the bond efficiently in this new set is desirable, even though it is excess from a narrower point of view, and may make the straightforward "seven eights are fifty-six" over-learned. So also a pupil's work with 24, 34, 44, etc., $+ 9$ may react to give what would be excess practice from the point of view of $4 + 9$ alone; his work in estimating approximate quotient figures in long division may give excess practice on the division tables. There are many such cases. Even adding the 5 and 7 in

5 7

$- + -$ is not quite the same task as adding 5 and 7, undisturbed

12 12

by the fact that they are twelfths. We know far too little about the amount of practice needed to adapt arithmetical bonds to efficient operation in these more complicated conditions to estimate even approximately the allowances to be made. But some allowance, and often a rather large allowance, must be made.

The second is the case where the computation in general should be made very easy and sure for the pupil except for some one new element that is being learned. For example, in teaching the meaning and uses of "Averages" and of uneven division, we may deliberately use 2, 3 and 4 as divisors rather than 7 and 9, so as to let the pupil's energy all be spent in learning the new facts, and so that the fraction in the quotient may be something easily understood, real, and significant. In teaching the addition of mixed numbers, we may use, in the early steps, cases like

$11\frac{1}{2}$	$79\frac{1}{2}$
$13\frac{1}{2}$	$98\frac{1}{2}$
24	rather than cases like 67

so as to save attention for the new process itself. In cancellation, we may give excess practice to divisions by 2, 3, 4, and 5 in order to make the transfer to the new habit of considering two numbers together from the point of view of their divisibility by some number. In introducing trade discount, we may give excess practice on "5% of" and "10% of" deliberately, so that the meaning of discount may not be obscured by difficulties in the computation itself. Excess practice on, and overlearning of, certain bonds in thus very often justifiable.

The third case concerns bonds whose importance for practical uses in life or as notable facilitators of other bonds is so great that they may profitably be brought to a greater strength than 199 right out of 200 in 2 sec. or less, or be brought to that degree of strength very early. Examples of bonds of such special practical use are the subtractions from 10, $\frac{1}{2} + \frac{1}{2}$, $\frac{1}{2} + \frac{1}{4}$, $\frac{1}{2}$ of 60, $\frac{1}{4}$ of 60, and the fractional parts of 12 and of \$1.00. Examples of notable facilitating bonds are "ten 10's = 100," "ten 100's = 1000," additions like $2 + 2$, $3 + 3$, and $4 + 4$, and all the multiplication tables to 9×9 .

In consideration of these three modifying cases or principles, a volume could well be written concerning just how much practice to give to each bond, in each of the types of complex situations where it has to operate. There is evidently need for much experimentation to expose the facts, and for much sagacity and inventiveness in making sure of effective learning without wasteful overlearning.

The facts of primary importance are:

- (1) The text-book or other instrument of instruction which is a teacher's general guide may give far too little practice on certain bonds.
- (2) It may divide the practice given in ways that are apparently unjustifiable.
- (3) The teacher needs therefore to know how much practice it does give, where to supplement it, and what to omit.

- (4) The omissions, on grounds of apparent excess practice, should be made only after careful consideration of the third principle described above.
- (5) The amount of practice should always be considered in the light of its interest and appeal to the pupil's tendency to work with full power and zeal. Mere repetition of bonds when the learner does not care whether he is improving is rarely justifiable on any grounds.
- (6) Practice that is actually in excess is not a very grave defect if it is enjoyed and improves the pupil's attitude toward arithmetic. Not much time is lost; a hundred practices for each of a thousand bonds after mastery to 199 in 200 at 2 seconds will use up less than 60 hours, or 15 hours per year in grades III to VI.
- (7) By the proper division of practice amongst bonds, the arrangement of learning so that each bond helps the others, the adroit shifting of practice of a bond to each new type of situation requiring it to operate under changed conditions, and the elimination of excess practice where nothing substantial is gained, notable improvements over the past hit-and-miss customs may be expected.
- (8) Unless the material for practice is adequate, well-balanced and sufficiently motivated, the teacher must keep close account of the learning of pupils. Otherwise disastrous under-learning of many bonds is almost sure to occur and retard the pupil's development.

INTELLIGENCE AND ITS MEASUREMENT: A SYMPOSIUM*

VIII. By V. A. C. HENMON,
University of Wisconsin.

1. *The nature and measurement of intelligence.* Intelligence, in the ordinary acceptance of the term, has been defined by Lester F. Ward to be "intellect coupled with the product of its operation," or in other words, "intelligence is intellect *plus* knowledge." This certainly corresponds to the use of the term in pedagogical and sociological theory and in practical life. The intelligent man is the well informed man and one who is capable of readily appropriating information or knowledge. Without doing violence to language the psychologist can not give it any other technical significance. Intelligence, then, involves two factors;—the capacity for knowledge and knowledge possessed. The untutored savage or barbarian may have high intellectual capacity, but without knowledge we should not ordinarily call him an intelligent man. We could best only say that he had a fine mind, a fine intellect, high intellectual power, or else that he possessed a high native intelligence. But native intelligence from the point of view of this definition is a misnomer, as is also the customary distinction between intelligence tests and achievement tests. Nothing but confusion results from using the term intelligence to designate the level reached in the acquisition of knowledge or mental age, in which sense it is something which can be developed and trained, and using the same term to designate intellectual capacity or mental alertness which is probably given once for all with one's constitution and can not be materially changed by anything we do. Achievement tests while measures of knowledge possessed are in turn indices of capacity for knowledge, though valueless for comparative purposes of crude, uninstructed intellect, only because there has been no guarantee of equality of opportunity. Intelligence tests, on the other hand, as at present constituted, are a mixture in varying proportions of tests of knowledge, for example, the constantly recurring tests of arithmetical reasoning and of information, with those whose novelty is such that specific past experiences function much less readily and there is an approximation toward relative equality of opportunity. In any case, both groups of tests are measures of both native and acquired intel-

*Continued from the March, 1921, issue. In that issue appear statements by Doctors Thorndike, Terman, Freeman, Pintner, Colvin, Ruml and Pressey.

ligence with differences in emphasis. The high correlations between intelligence tests and achievement tests and the high correlations between intelligence tests and teachers' estimates or scholastic standings indicate their essential similarity. Teachers' estimates and school marks, which are used as measures of reliability and validity of intelligence scales, in default of any absolute standard of reference, are based on what children know and on their capacity to know.

Intelligence ought to retain this broad significance. To define it with Stern as "a general capacity of an individual consciously to adjust his thinking to new requirements," which is substantially James' definition of reasoning, or with Meumann as a general capacity for "independence, originality and productiveness in thinking" is to narrow its meaning arbitrarily and to neglect its compound nature. The scholarly or erudite man who has merely acquired the knowledge created by others may not represent as high a degree of intelligence as one who is independent, original and productive in his thinking, but we should scarcely say that he is unintelligent. Intelligence is indicated by the capacity to appropriate truth and fact as well as by the capacity to discover them.

What we want of course, is a measure of crude intellect or mental alertness as such, a scale for intellect not a scale for intelligence. Any test which requires mental processes, be they either on the sensory-motor level or on higher levels, which measures the sensitiveness, responsiveness, and retentiveness of the nervous system, is a test of intellectual capacity. It is only because of the absence of an absolute standard of reference, the inability to discount the effect of past experiences, and the narrow conception of the meaning of intelligence, that the tests of higher mental processes appear to be better measures of intellect, and not because they involve intellectual capacity any more evidently. The essential thing in a test of pure intellect is to secure such novelty as to reduce inequality in opportunity and training to a minimum, and such complexity as to reveal differences clearly and unambiguously. It is not a matter of the kind of mental process involved primarily. A properly constituted test of intellect ought, therefore, to take a wide sampling of so-called lower as well as higher processes and the relative weights to be assigned to each element determined. The tests must be extended to measure the capacity to learn in other directions than the ability

to use words, signs, and symbols. The so-called general intelligence tests are not general intelligence tests at all but tests of the special intelligence upon which the school puts a premium. The extension of the term intelligence to include what Thorndike has called mechanical intelligence, the ability to manipulate things, and social intelligence, the ability to manage men, means that we need not one, but several scales of intellect from which we can secure an intellect profile. We have been gradually evolving a type of test for abstract intellect that is of use in classification and prognosis in school, but is of no convincing value as a measure of general intelligence as such, nor of demonstrated value in vocational guidance and direction.

2. *Next steps in research.* The construction of group tests proceeds apace. Many of them are uncritically and hastily assembled, hurried into print without norms or standards to satisfy the demand of some publishing company, and often merely rearrangements of familiar material, selected on the basis of the author's opinion of the merits of the component tests. While experimentation in test construction is desirable, critical evaluation of the more important group tests is far in arrears and much needed. General mental age determinations from existing tests are seriously open to question; the intelligence quotient derivatives involve questionable assumptions as to the nature of mental growth; the relatively narrow range of functions measured in many teams of tests make them invalid as measures of general ability; the assignment of equal weight to each test and to each item in the individual tests is surely wrong. These matters need careful study with the better known and better standardized tests.

The correlations between the scores in many of the group scales as a whole and such standards of reference as we possess do not differ very materially. The individual tests, however, differ among themselves considerably in diagnostic and prognostic value, in susceptibility to practice, in interest to the one tested, in ease of administration and scoring, in relative emphasis on previous training and capacity for meeting relatively new situations. What is needed clearly is more accurate scaling of the items within the individual tests, the reduction in number or elimination of tests that test apparently the same or similar functions, the application of the method of partial correlations to determine causal relationships and the correct weights to be assigned to the individual tests in the

scale as a whole. Besides the refinement in the tests of abstract intellect themselves we need also to determine the importance of various character traits which apart from intellect as such make for success in the tests. Until these things are done we shall be in constant danger of misusing the results of tests for classification and prognosis even in schools. We are still far from being able to give the significance and meaning of a test score.

Besides the need for critical evaluation, standardization and interpretation of tests of abstract intellect stands the need for tests of ability to manipulate things and the ability to deal with men, leadership and adaptability in social relationships to complete the intellect profile suggested above.

IX. By JOSEPH PETERSON,
George Peabody College for Teachers.

1. *The nature and measurement of intelligence.* Intelligence seems to be a biological mechanism by which the effects of a complexity of stimuli are brought together and given a somewhat unified effect in behavior. It is a mechanism for adjustment and control, and is operated by internal as well as by external stimuli. The degree of a person's intelligence increases with his range of receptivity to stimuli and the consistency of his organization of responses to them. In certain higher learning processes the bright subject more quickly and more systematically enlarges the range of his cues or 'guide-posts,' and consequently more readily masters the situation than does the dull one, who must come back again and again to the more immediate elements. The latter cannot keep in mind the complexity of the situation. This difference is found to hold for children of very different years but equal mental ages, contrary to the results of a well-controlled experiment by Woodrow, one, however, the results of which seem to me to have been generalized too much by him and by some other writers. A dull and a bright child of equal mental age have gone through equal mental-age levels, but the bright child seems to have made the same acquirements in the lesser time.

One's emotionality, responsiveness to abstract symbols, degree of energy and of perseverance, etc., certainly condition one's intelligence. We cannot analyze such differences here, but can only suggest that equal mental ages by any test do not indicate just equal

intelligence for all situations. We are probably too prone in our practical work of testing to assume equality of intelligence in such cases, since our attention is taken up with the one kind of reaction most susceptible to testing. General intelligence, if it is a reality at all, is probably not a separate or constant factor, but a composite of many different abilities, and probably means different things in unlike situations, as different abilities are stressed. Such factors as energy and perseverance, degree of disturbance by emotions and self-consciousness, and many others that play their rôles in one's success in life, have not yet been successfully brought into the field of measurement by tests, especially by group tests.

We have moved a long way from testing intelligence by specific sensory discriminations, reaction time, etc., and have rightly come to emphasize constructive imagination, quickness of the perception of relations, and other higher mental functions; but so far as the production of group intelligence tests that will measure abilities making for success in many phases of practical life is concerned, we have not yet arrived at an entirely satisfactory solution. Undoubtedly we shall need different tests for different purposes, though general intelligence tests seem to be the most promising on the whole; but our chief problem is to present the suitable problematic situation. Reactions with a pencil to highly abstract symbols certainly constitute an inappropriate type of test for an adult, let us say, who has spent years developing specific habits of other kinds. Reactions to pictures are far from being suitable substitutes for the more concrete situations demanded in the case of illiterates, although with proper precautions picture tests can be made very serviceable. Other objective materials are not yet very suitable for group tests.

2. *The next step in group mental testing.* The "next step" will probably not be a radical departure from present methods. Some refinements in method that seem to me probable and desirable are:

1. More attention to the equality of score points and to the treatment of errors. At present the score value of different test-units is largely a matter of personal judgment. Some 'scales' increase in difficulty more or less, uniformly in successive units, but usually even such increases are not accurately evaluated. In the most widely used scales units of increasing difficulty are found in some of the tests, while in other tests the units are of approximately equal difficulty. Even though fairly satisfactory results are thus

obtained, especially in the selection of individuals of extreme abilities, it would be sanguine to hope for much development in the science by such methods. No one knows just how much we fail with individuals of more nearly median ability, who differ from one another by smaller amounts; and to find this out requires more perfection in the criteria by which we usually evaluate tests. The development of methods for obtaining more reliable criteria is much needed, otherwise one test is correlated with another and the other one is in turn evaluated by it.

In the scoring of errors Thurstone has made a good start, but his method presupposes linear regressions and approximate equality of score units. An illustration may show its value. If we use the total score in the Pressey Cross-Out Test, Schedule E, as criterion—the only one available—and seek by Thurstone's formula to find the best scoring value of C in the equation $S = R + CW$, in which S stands for score, R for rights and W for wrongs, we get from tests of 53 12-year-old unselected children in Nashville schools, $C = -.44, .69, -.24$, and $.21$, for tests 1, 2, 3, and 4, respectively. Using these values of C , we get the following correlations of these four tests, in order, with the criterion: $rI(R + CW) = .61, .96, .53$, and $.63$. The corresponding values of r between total scores and rights (neglecting errors as was done in the derivation of the standards) are $.55, .71, .48$, and $.69$. This example shows that different parts of the same test may have, and probably should have, different score methods. These the authors of future tests should determine, and they should prepare tables to facilitate rapid scoring. Such treatment of group tests, following careful evaluation of score units, would greatly increase their usefulness, but it requires reasonably reliable criteria.

2. More thorough-going analysis of correlation coefficients. In the first place, we must not forget that $r = \sqrt{b_1 b_2}$, and that therefore it may be very deceiving in itself. There are also many factors making for non-linearity of regression lines that need careful study, many of them being due to such imperfections in the test itself as improper time allowance, improper gradation as to difficulty, and poor scoring methods. Formulae for non-linear relationships seem to me to have very little practical value for interpretation. We have found it more profitable to make distribution tables of scores and to study the actual curves through the means of rows and columns than to use these formulae for anything but merely to show the de-

gree of relationship. Such an analysis may, for instance, show that a test has significance only for certain degrees of mental ability, as at extremes, a condition that the correlation ratio alone could not reveal.

It should be noted also that correlations have very little significance for comparisons with others unless the degree of heterogeneity of the group tested is indicated. A low correlation may in some cases be better than a high one in others. There is little significance in the value of r itself. Partial and multiple correlation methods will doubtless be used in increasing measure both in the selection of suitable tests and in the analysis of results. Another illustration from the test of twelve-year-old children may show this better than will any other method. The four tests—counting rights only—showed inter-correlations as follows: $r_{12} = .33$, $r_{13} = .31$, $r_{14} = .20$, $r_{23} = .26$, $r_{34} = .54$. The partial coefficients of the second order are $r_{12.23} = .22$, $r_{13.24} = .22$, $r_{14.23} = .02$, $r_{24.13} = .49$, $r_{34.12} = .03$. It is seen that tests 1 and 4 have nothing in common not included in 2 and 3, and that 3 and 4 also test different functions but for the effects of 1 and 2. On the other hand, 2 and 4 have much in common not included in the others. (This analysis is based on Pressey's method of scoring, however.) On the basis of facts such as these it is possible to select a battery of tests that measure widely different functions, and thus to get better general intelligence tests as well as valuable data on the interrelation of mental functions. A better scoring method in this case might reveal different relations from those here indicated, but we are interested only in the method of procedure.

In the April number of this journal I have indicated some important problems relative to the time element in tests.

X. By L. L. THURSTONE, ✓
Carnegie Institute of Technology.

What is meant by intelligence?

Intelligence as judged in every-day life contains at least three psychologically differentiable components: a) the capacity to inhibit an instinctive adjustment, b) the capacity to redefine the inhibited instinctive adjustment in the light of imaginably experienced trial

and error, c) the volitional capacity to realize the modified instinctive adjustment into overt behavior to the advantage of the individual as a social animal.

The inhibition of an instinctive adjustment involves the substitution of an overtly passive deliberative attitude against environmental, social, and instinctive pressure. The degree of intelligence is partly measurable by the relatively early inhibition of the instinctive adjustment at a stage when it is not yet defined.¹ Conceptual thinking is rendered possible by the inhibition of the instinctive adjustment at a very early stage of its formation. A concept is an unfinished act inhibited before it has become personally concrete and it therefore has more extension and less intention than the idea which is an act inhibited after it has become motorially more defined. The capacity to redefine the inhibited instinctive adjustment in the light of imaginably experienced trial and error is essentially the analytical element of intelligent work. In routinized intelligent work the mental trial and error experience may be rationally controlled, but in unroutinized intelligent work the mental trial and error experience has its causal antecedents in loose affective and conative analogies. These subconscious analogies of intelligent work constitute alternatives of less resistance at a preconscious stage of the unfinished, undefined and inhibited instinctive adjustment. Analytical capacity may be thought of as flexibility of affective and conative associations which precede the cognitive phase in the definition of the stimulus into the response.

These three phases of intelligent behavior may be labelled for convenience a) inhibitive capacity; b) analytical capacity, and c) perseverance. One may possess the inhibitive capacity without keen powers of analysis. This leads occasionally to an appearance of intelligence by silence. Such a mind is often capable of recognizing and adopting the successful analyses of other minds without being able itself to produce the analyses.² The inhibitory capacity for the early, abstract, impersonal stage of instinctive adjustment does not necessarily coexist with inhibitory capacity for the later more defined and personally significant stage of the instinctive adjustment.

¹By "early" I do not refer to childhood or the education of youth. I mean by "early" the incipient response as it is being consciously particularized in every moment of normal waking life.

²When such a mind is by superior social traits adapted for administrative work it will lean on competent assistants for analyses. This type of mind is common in industry and in a deplorably exaggerated form in politics. It was described even by Machiavelli in "The Prince."

Hence a man may possess inhibitory capacity for abstract impersonal thinking and lack self-control for those instinctive adjustments which have reached uninhibited a personally and socially significant and overtly defined stage. A man may possess considerable inhibitory and analytical capacity and still be socially quite ineffective because of insufficient volitional drive to realize his intelligence into consistent action. Such a mind is not infrequently the subject of ridicule from those who would justify their own mental worth in spite of modest analytical powers.

Analytical capacity concerns the preconscious and early conscious stage in the definition of instinctive adjustment. It implies the capacity to inhibit instinctive pressure at that stage so as to allow flexibility of definition of the response. This has been picturesquely called by Knowlson free interplay with the subconscious.³ It is only natural to expect that such inhibitory capacity should appear when the instinctive drives are relatively weak. Hence superior analytical ability is not infrequently associated with inferior instinctive energy in its nutritional, sexual, gregarious, and self-assertative forms. When such an individual discovers his inferior instinct equipment he may center his self-satisfaction on his analytical powers as a compensation. Such individuals need the assistance of other volitionally stronger minds in order to make their labors socially effective.

When the analytical capacity is associated in the same mind with a fair degree of instinctive and volitional energy we have the bright person in practical life. There is a fundamental distinction between bright and profound intelligence. The mind which possesses a fair amount of inhibitive, analytical, and volitional capacity will in general aim for effectiveness in the immediate present. Such a mind is bright. When, however, a mind is more generously endowed in all three of these respects or when it is unbalanced by reduced volitional capacity, it deals with the earlier and more abstract relations in its adjustments. Such a mind is profound and its analyses concern temporarily remote ends. I am quite sure that our intelligence tests in which the candidate races against time for a few minutes do not measure adequately the more inhibited and deliberative profound type of intelligence. It does fair justice to the more frequent well balanced bright mind.

³"Originality," by T. S. Knowlson.

I should summarize my definition of intelligence as follows: Intelligence is the capacity to inhibit instinctive behavior in an unfinished stage of its formation and to modify it at that stage by means of an imaginal stimulus which is relatively remote from that which is immediately and perceptually present. The imaginal stimulus in intelligent behavior is of preconscious origin. It becomes conscious as an alternative to be controlled, accepted or rejected rationally, but its origin is a preconscious, irrational, uncontrolled association of affective and conative similars. This definition is consistent with that of Baldwin in his discussion of meaning in which he says: "It is in the passage from the bare recognition of each item presented as being just what it is, to its treatment as being in some sense not what it is, but what it may become or be used as, that psychic meanings as such arise." Bergson says that "intelligence, considered in what seems to be its original feature, is the faculty of manufacturing artificial objects, especially tools to make tools, and of indefinitely varying the manufacture."

There are two features in intelligent behavior that I have tried to emphasize, namely, that it implies inhibition and that it consists largely in rendering conscious an unfinished act. The sooner the act becomes conscious, the more crude and unfinished it is when it becomes conscious, the greater will be the range of possible overt behavior into which it may become particularized, and the higher will be the intelligence of that particular moment. I have previously described this point of view with reference to the momentary psychosis.⁴ The element of inhibition in intelligent behavior may be partly justified if thought of in relation to its various ramifications such as the inhibitory functions of the cortex, the recession of the stimulus in intelligent behavior,⁵ the frequently drawn contrast between intelligent behavior and instinct- or habit-determined behavior, the modification of instinctive protopathic behavior into intelligent epicritic behavior,⁶ and the analytical intelligence of the inhibited introvert as contrasted with the analytical impatience of the extrovert.

By what means can intelligence be best measured by group tests?
This question can be given at least three different interpretations.

⁴"The Anticipatory Aspect of Consciousness," *Journal of Philosophy, Psychology and Scientific Method*, Oct. 9, 1919.

⁵Holt: "The Freudian Wish and Its Place in Ethics."

⁶Rivers: "Instinct and the Unconscious."

We may set out to measure pure intelligence as a trait which is only partly responsible for social success. In this case we shall probably have no criterion by which to judge the test. We may set out to measure brightness using as our criterion the estimates of avowedly competent judges of brightness and dullness with the necessary admixture of volitional and emotional ingredients. We may set out to measure ability to do a specified practical task for which we assume that intelligence is an indirect partial requisite. This latter type of mental measurement is the customary one. Our criterion is the relative success of the candidate in school, college, salesmanship, clerical work, or what not. Any test procedure which has diagnostic value with reference to a specified criterion is acceptable. It is much more difficult to measure intelligence as such and I am not sure that it has ever been done. For administrative purposes, we do not need to measure pure intelligence. In fact our diagnoses are probably more effective if we test other significant factors in our tests, although it would be scientifically interesting to be able to tease out the intellectual, volitional, affective, social, heredity, and experiential elements.

It would be well if we could draw a sharp line of distinction between service and research in mental test work. These two purposes are often confused. The results of research may be summarized for direct application in some mental test service. Some research data may be obtained from service records but the reverse is rarely successful. Some mental testers are justifying their use of a test in test service simply because they have norms for it. This should never be done. No test should ever be used for any kind of service unless it is known what it is that the test diagnoses. This necessitates a criterion for every test. If the test correlates well with a criterion such as chronological age, it can be used to determine age. But a test may be good for one criterion and poor for another criterion. We should never talk about a "good" test without telling what it is that the test is good for—namely, the criterion. Another relatively common mental test illusion is that if a new test has a high correlation with the Army Alpha, and if the Army Alpha is good for some specified criterion, the new test is forthwith assumed to be good for that criterion too. This does not follow. The only safe way is to take the trouble to determine the diagnostic value

of the new test without inferring it through another test as an intermediary.⁷

It has always seemed to me rather peculiar that we measure children's intelligence in terms of their chronological age. I should like to suggest that possibly the underlying assumptions would be less troublesome if we stated the intelligence of an eight-year old in terms of his percentile rank with reference to other eight-year olds, or in terms of the standard deviation of test scores for eight-year-old children. From such a direct rating one could of course readily ascertain that chronological age the mean intelligence test score of which any given child attains.

What are the most crucial next steps in mental test research? Now that we have accumulated enough norms and correlation coefficients to make one dizzy it is about time that we begin to formulate some mental test principles. Our literature abounds with instructions for giving particular tests and with statistical data of administrative interest. But we have so far no science, no principles, no psychology in the mental test literature. We have done well with our empirical cut-and-dry methods in mental test work. We can do better if we rationalize our findings in this field. Unfortunately there are relatively few mental testers who are interested in deriving psychological fundamentals from mental tests.

I should like to see another line of mental test work opened up, namely, the diagnosis of the volitional and emotional characteristics which determine our character traits. Intelligence is only one of the elements in mentality and it has been overworked because of being accessible to measurement. We should investigate the possibility of diagnosing character traits by some new kinds of mental test, self-analysis forms, and other procedures. Who knows but that individual differences in the various characteristics of our reflexes may be diagnostic of character traits? Why is it that executives are in general taller and heavier than other people with similar habits of life? Why is it that we guess right more often than wrong about the character traits of a stranger by merely looking at his face? These questions should be rationalized as science. What are the possible schemes by which we may successfully classify character traits? What types of mentality are produced by the combination of intelligence and different emotional and volitional traits?

⁷Yule: "Introduction to the Theory of Statistics," p. 250.

By what physiological and psychological technique can we diagnose these types? What are the mental processes that distinguish special ability? Can these be taught to others? These are worthy problems for mental test psychologists.

It is high time that we quit justifying ourselves as psychologists by simply standardizing mental tests. If we attack the individual diagnosis of character traits as energetically as we have been giving group tests the results will be of far reaching psychological, educational and social significance.

XI. By HERBERT WOODROW,
University of Minnesota.

1. *The nature and measurement of intelligence.* In attempting to state the meaning of intelligence, I should say first that it is an acquiring-capacity. One's actual, present capacity for doing any particular thing is obviously dependent upon his training and learning, or, in general, upon his past experience; whereas, in the measurement of intelligence, the attempt is always made to minimize the effects of past experience by utilizing performances, the acquisition of the capacity for which is presumably favored equally by the past environments of all the individuals measured. When, for example, we use a child's present capacity to count backwards from twenty as a partial index of his intelligence, the assumption is, clearly, that the past environments of all children have been equally favorable for the acquisition of this ability. If this ability was one which could be readily taught, and, further, one which was taught some children but not taught others, it would be valueless as an index of intelligence. The present capacity to count is a measure of intelligence only in so far as it is an index of the capacity to learn to count, that is, the capacity to acquire the capacity to count. Intelligence, then, is the capacity to acquire capacity.

Some psychologists, no doubt, would urge the desirability of defining intelligence as a capacity which is fixed at an early age and is largely innate. These, while conceding that intelligence is a thing that *grows*, might argue that the capacity to acquire is, itself, largely acquired. They might point out that, at the age of fifteen,

Poww

for example, one has the capacity to acquire abilities that could not possibly be acquired at the age of six, and that the difference in the capacity to acquire at these two ages was largely due to what had been learned in the years between them. I think this argument is sufficiently answered, however, by regarding the acquisitions between any two ages as progressive steps in the acquiring of an ability, that is, as part of the learning of whatever is learned easier as a result of them. I should, therefore, incline to the view that it is not a necessary part of the *definition* of intelligence, but simply an established fact, that intelligence, except for being a *growing* thing, is fixed partly by heredity and partly by environmental factors acting before the age of five. I hold, further, the view which I proposed in this Journal some years ago (1917) as an experimental conclusion, that the growth of intelligence is neither produced, nor appreciably accelerated, by learning.

Now, if intelligence is to be defined as a capacity to acquire capacity, it is imperative to define the acquired capacity. I believe this capacity may be correctly described (using neither purely mental nor purely behavioristic terms) as the capacity for such mental activity as is most effective in bringing about success, or, in slightly different phraseology, as the capacity for success in so far as success is dependent upon mental processes, either past or present. By success, is here meant, not success in any one line of endeavor, but a theoretical success, namely, success in all those innumerable performances which may be regarded as desirable. The capacity for such a theoretical success would have as its practical result success in coping with any situation in which the individual might find himself.

This conception of intelligence implies that mind is an instrument for the securing of desirable ends. It implies that between mind and body there exists a relationship—a fact which I believe no living psychologist denies—but does specify the ultimate nature of the relationship. Intelligence cannot be defined in purely mental terms, because the capacity for acquiring valuable modes of mental functioning is itself not mental; it is certainly, in part, a matter of the condition of the brain. Moreover, success can be attained only as the immediate result of behavior. All “mental measurements,” even those concerned with “intensity of sensation,” are primarily measurements or descriptions of behavior under measured or described environmental conditions. On the other hand, intelligence cannot

be satisfactorily described in purely behavioristic terms, because (I mention only one reason, one which is persistently overlooked by behaviorists) while the degree to which behavior is intelligent is simply the degree to which it attains success, success has no real meaning except by reference to some want, desire, intention, plan, or purpose.

From the above point of view, the problem of the relative desirability of different kinds of testing materials is essentially the problem of the relative value of the different kinds of mental processes in securing "success in general." The existing data show that many of the "simpler" mental processes are of value, as well as all, or nearly all, of the more complex ones. In general, however, the more complex processes are much the more valuable. Fine pitch discrimination, for example, while by no means useless, is of far less value than good attention and correct reasoning. From these considerations, it follows that any accurate measurement of intelligence, whether by individual or group tests, must call into play both the more complex mental processes and the simpler ones, but should invoke each, or weight the measurement of each, in exact proportion to its usefulness in securing "success in general." Since, as a rule, the more complex processes are the more useful, they should, as a rule, be given the greater weight.

Now the relative usefulness of the different forms of mental activity varies with sex, social status, stage of civilization and, most important of all, with degree of intelligence and therefore with *age*. The best tests, the tests which go farthest in determining the outcome of the measurement of intelligence, will differ with the age of the children measured, and, in general, will be tests of the simpler mental functions in young or unintelligent children and tests of the more complex functions in older or more intelligent ones. To use mainly tests of simple mental functions at the early ages, and mainly tests of more complex ones at the later ages, is simply one way, though probably the most convenient one, of giving greater weight to the measurement of simple mental functions in the early stages of the scale. However, even a scale designed solely for children of some specified early age will involve the measurement of capacities which, though perhaps all comparatively simple, will vary to some degree in complexity and in value; and even such a scale should, therefore, in accordance with the general rule stated above, give weight to the

measurements of different capacities in exact proportion to their usefulness.

2. *Next steps in research.* While I am unable to outline any "next steps" which I should call "crucial," there are numerous desirable ones. I should urge, in the first place, the desirability of repeating again and again the steps already taken. Investigators who have worked extensively in this field have, after all, been few; and the very fact that they have done so well entitles them to successors who can devote years of patient toil and the necessary originality to the devising of new methods of measurement; to a careful analysis of the mental processes involved in each test, together with the study of individual variations in the same; to the study of the relative weight of different tests as measurements of intelligence; to establishing norms, not from data furnished by untrained people, distorted by the "personal opinion" and "benefit of the doubt" methods of scoring, but from data obtained in a scientifically reliable manner; and, in short, to devising tests with a better logic behind them, both from the viewpoint of psychological theory and the mathematical theory of measurement.

More information is needed on every point connected with brightness, or relative intelligence, in distinction from absolute intelligence. No doubt further data will soon be forthcoming on individual variations in brightness with chronological age, with anatomical age, with disease, with harmones, with learning and with various environmental factors. In the measurement of brightness, even more than in the measurement of intelligence, we need further work in the logic of method. For studying the relation of brightness to anatomical age, we need, and we shall have, a much more adequate scale for the measurement of anatomical age than any now in existence.

XII. By W. F. DEARBORN,
Harvard University.

1. The commonly accepted definition of intelligence such as the capacity to learn or to profit by experience evidently involves a description of the nature of the individual and his reactions to the environment, which, I believe, can now be adequately given in terms of the current "objective" psychology. The older conceptions of per-

ception, association, imitation, and reasoning are described concretely in terms of "unconditioned" and "conditioned" response, and learning, as the development of intelligence, appears as a continuous process from the establishment of the first conditioned response to the "discriminating perception" called reasoning.

2. Theoretically, it would follow that measurement of the actual progress of representative learning would furnish the best test of intelligence. For practical reasons most tests now in common use are not tests of the capacity to learn, but are tests of what has been learned. The assumption is made that if one samples the results of learning in matters where all the individuals tested have had an equal chance at learning, he may arrive at an estimate of the capacity to learn. But, since it is difficult to find even simple experiences which are common to all individuals of a given age period, actually, again, one tries by sampling a large range of fairly common experiences to strike an "average" which, despite the fact that a given individual may have missed this or that experience, will still be representative of the individual's learning. The shortcomings in the working out of these assumptions are generally recognized. An extension of this method would still seem desirable and the development of methods of group testing will make possible the more rapid trying out and evaluating of the tests that are adapted to group use. The best of these should be used to extend and make more representative the individual tests now in use. A wider sampling of experiences is clearly necessary in the upper age periods. In addition, individual tests involving actual learning rather than of the results of learning are needed.

3. The methods of determining the mental "age" or status of an individual need revision. The mental age is the resultant of at least three factors, native intelligence, physiological maturity and "environment." An analysis of these factors may be secured by the repetition of mental and physiological tests during a period of years on the same individuals. Assuming for a given group of individuals a constancy in the first and last factors the variations in the second factor, particularly during the adolescent period, will certainly show some variability in the individual intelligence quotients.

4. Finally, the whole question of the unit of measurement assumed in mental age and intelligence quotient needs careful canvassing. Pearson has stated, as a result of his study of the somewhat

limited material of Jaederholm, that a year's mental growth is for a number of years practically constant and equal to the standard deviation of the groups. The conclusion has been drawn from Bober-tag's study of the differing amounts of overlapping in the 7th and 8th years as compared with the 11 and 12 year olds that "a year of mental age at the younger ages means a bigger change than it does at the higher ages" (Woodrow). Both statements assume a normal distribution and a fixed variability. The latter assumption is certainly open to question. And, suppose that growth is comparable to the practice curves of learning, where a small gain towards the end of the learning seems harder to make than a larger gain at the start, is this more than saying that your arbitrary units of measurement are changing? Careful use of the results of the group methods of testing should give data for the attacking of these problems.

XIII. By M. E. HAGGERTY,

University of Minnesota.

1. *Intelligence and its measurement.* In responding to this request to participate in the symposium on intelligence I take it that what is desired is a more or less dogmatic confession of faith rather than a systematic effort to support by evidence and argument the beliefs expressed.

1. In my thinking the word intelligence does not denote a single mental process capable of exact analytic definition. It is a practical concept of connoting a group of complex mental processes traditionally defined in systematic psychologies as sensation, perception, association, memory, imagination, discrimination, judgment and reasoning. The nearest equivalent to this working concept is James' "differentia of reasoning" as "the ability to deal with novel data" and Dewey's well known "forked-road" theory. This limitation of the concept of intelligence to the "novel-situation" category, so useful for theoretical purposes, however, seems to me too thin and lean for practical purposes of controlling behavior. I prefer to enrich the concept with all the fatness afforded by the inclusion of the more elementary and subsidiary processes instrumental to the crucial act of thinking or reasoning. For the most part, I would

exclude from the concept, emotions, instincts, will-activities and so called character traits.

2. Intelligence is a dynamic concept. It is descriptive of behavior and not of static component parts of the "mind." When one conceives of the mind as a "state" or a "structure," the word intelligence becomes meaningless. The implication of *activity* is essential in the concept of intelligence.

3. Intelligence is native and peculiar to an individual. In this sense it is an instinct defined as an inherited or congenital characteristic. It may be conceived of as akin to, if not dependent on, or a phase of, the native quality of the physical organism, the brain and nerves. Because of the original differences in intelligence or native capacity the same environment affects differently different persons. As a result two persons subjected to the same environment, say the formal procedures of the elementary school, emerge with different mental content, different habits, different interests. These eventuating differences are the criteria by which we may infer the quality of the native endowment of the individual. Often these differences among individuals are less important in themselves than they are for such inferential purposes. Present attainments are important in showing what an individual can now do. In so far as such attainments may be made the basis of inference as to native endowments they are of value for prognosis. Because of their superior prognostic value intelligence tests have a value which mere achievement tests do not have.

4. Intelligence is not merely a qualitative concept. It may be quantitatively considered. When a child has little of it he does poorly in school, and in cases of extreme deficiency he is largely uneducable. When a child has more intelligence he is capable of learning and individuals gifted with great intelligence may learn most rapidly. Similarly in actual life low intelligence means simple occupations and crude civilization. High intelligence means the possibility of efficiency and leadership in the more complex problems of civilized society.

5. Because intelligence is capable of quantitative considerations it is theoretically measurable. Measuring devices with a considerable validity have within recent years been made available in the form of individual and group scales and tests. These scales and tests vary greatly in value, but in increasing numbers they are

meeting essential statistical criteria and through the improvement of the technic of administration and interpretation they are becoming widely usable and useful.

The most satisfactory tests so far devised attempt to measure intelligence in its essential complexity. The earlier attempts to analyze the factors of intelligence into simple elements such as sensation, perception, memory, attention, etc., largely resulted in failure to secure useful tests. Over-analysis apparently distorts the essential phenomena to be measured and to follow this path of experimentation is to court futility. The practically useful test must evaluate the capacities of an individual in the essential complexity necessary to meet actual living situations.

6. The most satisfactory intelligence tests so far devised very largely involve the use of language. Of these the most dependable are tests of a knowledge of vocabulary, of general information, tests for facility in thinking logical relations such as opposites and analogies, for ability in the solution of arithmetical problems, in the completion of elliptical sentences, etc. It is possible and probable that such verbal tests will always be the best and most generally useful means of testing the higher ranges of intelligence. Such verbal tests, however, have their limitations in testing illiterate or partially illiterate adults and children, in testing children in the pre-school years, and in testing such individuals as may find the most adequate expression of their intelligence through manual or other non-verbal activities.

7. An intelligence test is not an end in itself. It is an instrument for diagnosis. It is intended to aid a capable and intelligent person to secure a more accurate determination of the intelligence of a human being than he could secure without such a test, to secure it in a shorter time and in general to enable such a capable and intelligent person to deal with human beings in a more discriminating fashion than he otherwise could do. The score on an intelligence test, however, is not to be regarded as a complete substitute for personal judgment and common sense on the part of the examiner. The analogy most expressive to my way of thinking is to compare the intelligence test in the hands of a psychologist or educator with a thermometer in the hands of a physician. In a similar fashion the test is an instrument for recording symptoms. Its results are one aid in diagnosis.

8. Where education pretends to be based on fundamental principles it must take account of the factor of intelligence. A measure of the intelligence of pupils must, therefore, be available to the teacher, the supervisor, the administrator and possibly to the pupil himself. The extension of intelligence examinations to the public schools under conditions which properly safeguard the interpretation of results seems, therefore, a legitimate movement. The time will probably come when all progressive schools will record the intelligence score of a pupil with the same care that it records his chronological age. For educational purposes, it is more important.

9. Students of experimental education must take account of the intelligence of experimental subjects as they have not done in the past. Most of the studies designed to compare one method of learning with another are deficient in this respect. In many cases the varying intelligence of the subjects is a more determinative factor in the results than is the method of learning used. Indifference to this factor has given us experimental results which are non-interpretable. Most of the experimental work in this field must be done over if it is to throw any real light on educational problems. The freedom with which certain experimentations have sailed along, indifferent to the factor of intelligence, casting ashore their partial findings, has done and is doing much harm.

10. When intelligence tests are used the results must be interpreted in the light of other data indicative of native capacity. In the case of school children such "other data" are school marks, teachers' estimates, school progress, etc.

11. Group intelligence examinations of thirty minutes' duration give a highly useful rating of school pupils. When supported by other data the results can be used as a basis for classification. The value of such a test will be increased by the repetition of the same or a similar test. Individual examinations should be employed in complicated cases.

12. Intelligence is not the only factor conducive to success. Industry, perseverance, loyalty, cheerfulness and other non-intelligence traits are important and in many cases determinative of success.

2. *Next steps in research.* The direction of experimental progress in dealing with intelligence problems seems to me about as follows:

a. The perfection in technique and statistical criteria of verbal tests for the ranges of ability where such verbal tests may be used. The general lines for work are clear; refinement of method is the end sought. Much remains to be done.

b. The development of non-verbal tests for young children, for illiterate and non-English reading children and adults, and for the examination of those special aspects of intelligence,—if any such exist—which are not properly measured by verbal tests. To some degree this means refinement of tests which have already demonstrated their usefulness. It probably also means the qualitative exploration of a new field.

c. The problem of special aptitudes is unsolved. Some evidence exists to show that skill in the analogies tests is prognostic of success in the formal freshman high school algebra. Possibly a number of our so-called general-intelligence tests have such specialized prognostic value. If so, the fact is worth experimental demonstration. It is highly probable that other aptitudes exist which none of our present tests detect. Such an aptitude might be a predisposition to learn telegraphy. It is altogether possible that such special aptitudes are sufficiently evident in childhood and youth to make their study and measurement of important educational significance. Certainly their quest is as inviting as many another which has issued fortunately.

d. Devices for securing the judgment of teachers on the capacities of pupils are undeveloped. "Rating scales" for pupils, for teachers, and for adults give promise of great usefulness if they can be redeemed from their present chaotic state and placed on a scientific basis. By their means we may be able to improve our evaluation of many non-intelligence traits not now subject to objective measurement.

e. The work of Dr. Downey on the "Will-Profile" suggests that the objective measurement of non-intelligence traits is possible. Probably nothing would better supplement our intelligence examinations than would the perfection of an objective measure of the so-called character-traits.

f. The establishment by some scientific organization enjoying public confidence of the essential criteria which satisfactory tests should meet would go far to reducing our discussions to a common language. The current proposal of the National Association of the Directors of Educational Research is a step in the right direction.

THE RELATIONSHIP BETWEEN EYE-PERCEPTION AND VOICE-RESPONSE IN READING

G. T. BUSWELL,
University of Chicago.

PROBLEM.

In oral reading the eye always moves along the line of print in advance of the voice, at times keeping very far in the lead and at other times very little in advance. A mature reader tends to maintain a comparatively wide average span between the eye and the voice, which at times may amount to the space occupied by seven or eight words. An immature reader, however, tends to keep the eye and voice very close together, in many cases not moving the eye from a word until the voice has pronounced it. Reading of this type becomes little more than a series of spoken words because there is no opportunity to anticipate the meaning in large units. An eye-voice span of considerable width is necessary in order that the reader may have an intelligent grasp of the material read, and that he may read it with good expression. If words are encountered which are spelled alike but pronounced differently, such as "read" (present tense) and "read" (past tense), the correct pronunciation and meaning cannot be determined in many cases until the eye has observed the context by looking ahead. The need for a wide eye-voice span is also emphasized when marks of punctuation are encountered. The failure of immature readers to respond to a question mark by a rising inflection of the voice is clear evidence that a narrow eye-voice span has kept them unaware of its presence until it is too late to modify their expression.

In order to determine more fully and accurately the nature of the eye-voice span an investigation was organized to cover a series of problems which were involved. This article presents a brief summary of the experiments, a complete report of which has been made in a separate monograph.¹

The experiments were grouped into three main divisions. In the first (a) an analysis was made of the eye-voice span showing the differences in the width of the span in the different grades and in the high school, and (b) the variations in the width of the span in different parts of the sentence. In both (a) and (b) the results were

¹Buswell, G. T. *An Experimental Study of the Eye-Voice Span in Reading*. "Supplementary Educational Monographs," No. 17. Chicago: The University of Chicago, 1920. Pp. xi + 105.

grouped to show separately the characteristics of pupils with mature and immature reading habits. Following this, (c) a comparison was made to show the relationship of the width of the eye-voice span to rate of reading, number of fixations per line, and regressive movements.

In the second division (a) a very detailed analysis of the eye-voice relationship was made, showing the exact position of the eye and voice at each fixation pause, and exhibiting the variations in the width of the eye-voice span in the reading of different parts of a passage by the same individual. These results were used (b) in an attempt to explain the cause of the occasional very long fixation-pauses which appear in reading records.

The third division of the investigation made use of a test device, consisting of a paragraph containing words spelled alike but pronounced differently, by which the eye-voice span in oral reading was studied in relation to the recognition of meaning in silent reading.

METHOD.

The apparatus and general method of the investigation were the same as used in a number of researches in the Chicago laboratory, and have been described in detail in a monograph by C. T. Gray.² Briefly stated, the method consisted of photographing a beam of light generated by an arc lamp, reflected first to the cornea of the eye from silvered glass mirrors, and then from the cornea through a lens to a moving film. The pencil of light changes its direction with each movement of the eye. As the subject reads a photograph is made on the moving film in the form of a sharply focused line. The shifts in this line record the movements of the eye. An electrically driven tuning fork, with a vibration rate of fifty per second, is mounted in the path of the beam of light in such a way that the light is intercepted at each vibration. These vibrations produce on the film a line of dots rather than a solid line, each dot representing a time of exactly one-fiftieth of a second. Since the film is moving continuously in the vertical plane, the record shows a vertical line of dots while the eye is fixated in a single position, and a short horizontal line when the eye is in motion in a horizontal or oblique

²Gray, C. T. *Types of Reading Ability as Exhibited Through Tests and Laboratory Experiments*. "Supplementary Educational Monographs," Vol. I, No. 5. Chicago: The University of Chicago, 1917. Pp. 83-90.

direction. By means of such a record the exact location of each fixation can be determined. In order to get a record of the voice the photograph was supplemented by a dictaphone record of the oral reading, taken for each subject at the time the photograph was made. By means of an electrical device the dictaphone and film records were synchronized, showing the relative position of the eye and voice at different places in the reading.

Photographs taken of the readings of 54 subjects selected as follows. Two good and two poor readers were selected from each of the elementary grades above the first, on the basis of scores made in W. S. Gray's Oral Reading Paragraphs. Three good and three poor readers were selected through the co-operation of the English department from each of the four high school classes. Six adult college students were selected at random and ranked into two groups, one better in reading than the other. The entire group of subjects, therefore, included twenty-four from the elementary school, twenty-four from the high school, and six college students, each grouping being made up of equal numbers of good and poor readers.

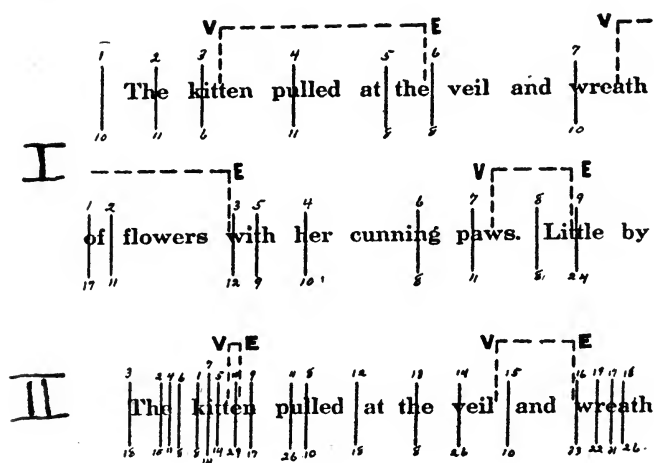


PLATE I.

First two lines show reading record of a good reader from the fourth grade.
Third line shows reading record of a poor reader from the second grade.

Plate I shows sections from two typical records which will serve to illustrate the method. The short vertical lines crossing the lines of print indicate the positions of eye-fixations. The serial numbers above the verticals indicate the order of the pauses; the numbers at

the lower end of the vertical lines give the duration of the fixation in fiftieths of a second. The brackets of broken lines indicate the relative positions of the eye and voice. For example, in the first line when the voice was pronouncing the word "kitten" the eye was fixated just after the word "the." The letters "V" and "E" show the corresponding positions of the voice and eye respectively, the width of the eye-voice span at this position being 17 letter-spaces. The third line of Plate I shows the record of a second grade reader with very immature reading habits. The eye-voice span for this subject is very narrow.

Analysis of Eye-Voice Relationship.

The first division of the investigation made an analysis of the eye-voice relationship and a comparison of the eye-voice span with other factors of the reading process. The results may be summarized as follows:

1. A positive correlation was shown between a wide eye-voice span and mature reading habits. The average span for the good readers is greater than that for the poor readers in every school grade. The average width of the span for the good readers in the elementary grades is greater than that of the poor readers by 58 per cent; for the subjects from the high school it is greater by 36 per cent.

2. Width of eye-voice span has a positive correlation with rate of reading, as shown in Table I which follows.

TABLE I.
RELATION OF EYE-VOICE SPAN READING TO RATE—ALL SUBJECTS.

Number of Subjects.	Rate of Words per Second.	Average Eye-Voice Span.
1.	0-0.9	3.4
2.	1-1.9	5.7
10.	2-2.9	11.3
29.	3-3.9	12.7
12.	4-4.9	16.5

3. A negative correlation is shown between width of eye-voice span and number of fixations per line. As the span increases in width the number of fixations per line decreases.

The three results above mentioned merely confirm what one would naturally expect to find true.

4. The development of the eye-voice span does not show a consistent increase from grade to grade, but is very irregular. On the whole there is a development in width of span throughout the grades, the high school average being greater than that for the elementary grades, and the adult average greater than that for the high school. However, some good readers from the elementary grades have a greater span than many of the high school subjects. The results show that some readers develop an eye-voice span by the end of the fifth grade which is greater than the average span for all the good readers from the high school.

5. Little correlation was found between width of eye-voice span and position in the line. However, when studied in relation to position in the sentence a marked correlation was discovered. Table II shows the average eye-voice span, in number of letter spaces, for positions at the beginning, within, and at the end of sentences.

TABLE II.
AVERAGE EYE-VOICE SPAN AT THE BEGINNING, WITHIN, AND AT THE END OF SENTENCE—ALL SUBJECTS.

Subjects.	Beginning of Sentence.	Within Sentence.	End of Sentence.
<i>Elementary subjects:</i>			
Good readers.....	16.1	14.8	8.4
Poor readers.....	8.6	9.5	7.9
Good and poor.....	12.3	12.1	8.1
<i>High-school subjects:</i>			
Good readers.....	16.6	16.2	13.3
Poor readers.....	12.9	11.0	10.5
Good and poor.....	14.7	13.6	11.9
<i>Adult subjects:</i>			
Good readers.....	23.5	18.6	14.0
Poor readers.....	17.7	10.8	11.3
Good and poor.....	20.6	14.7	12.6
<i>All subjects:</i>			
Good readers.....	18.7	16.5	11.9
Poor readers.....	13.1	10.4	9.9
Good and poor.....	15.9	13.4	10.9

An examination of these data reveal two conspicuous facts: first, that the width of the eye-voice span is different at various positions in a sentence, and second, that the good and poor readers do not exhibit these differences in the same fashion. The eye-voice span is found to be relatively wide at the beginning of a sentence and relatively narrow at the end.

The fact that the eye-voice span varies with the position in the sentence is of considerable significance. If the span varied only

with the position in the line, the determining factors would be only mechanical, and would be determined by the printed form of the selection. The control of the span in that case would be a matter of the mechanics of book construction, and would be independent of any teaching factor. But if the span varies with the position in the sentence, it is evident that the content of the meaning is recognized, and that the eye-voice span is determined by thought units rather than by printed line units.

For all three classes of subjects there is agreement among the readers in that a wide eye-voice span occurs at the beginning of a sentence. The situation at the beginning of a sentence is different from that of any other position. After one has started to read, the meaning of the thought covered will carry him along to some extent, and will enable him to anticipate what is coming. At the beginning of a sentence there is no sequence of words to give one the cue to the content of the new thought. The only way to get this is to look ahead until the meaning of the sentence is partially recognized, and the kind of vocal expression needed is made clear. The good readers recognize this need for a wider span at the outset and inhibit the voice reaction until the eye has gained a considerable lead. The poor readers in the grades above the elementary school have also learned this, but evidently those in the elementary school are not mature enough in reading to recognize any special difficulty at the beginning of a sentence. Instead of making a relatively longer span, they react to the situation by a relatively shorter one. They begin to read as soon as they see the sentence, and have not learned to inhibit their reading until the eye has taken in a larger unit of meaning. This difficulty could be easily corrected by a little training in class which would teach the pupils to wait before starting to read until they get a larger unit of thought.

The evidence of all subjects agrees that there is a shorter span at the end of a sentence. The good readers have a relatively shorter span than the poor readers. The explanation of this shorter span goes back again to the fact that the sentence is the large unit of meaning. When the eye reaches the end of this unit it modifies its movements according to the meaning recognized and the voice catches up before beginning the new thought. In order that the voice shall express the thought clearly, a pause is necessary at the end of the sentence. This pause gives the eye ample opportunity for a large

eye-voice span before it is time to commence the next sentence. A poor reader pays less attention to the sentence as a unit of meaning. This is especially true of younger children who are very immature readers. For them the whole process is a more or less monotonous repetition of words as they are encountered. The eye moves along at a regular rate and the voice follows. The end of a sentence creates no special disturbance, for it is passed over with little attention. Consequently there is little change in the eye-voice span. The data for the poor readers from the elementary school would seem to indicate that some such situation exists. There is little variation in the width of the span for any position in the sentence. If the variation in the eye-voice span at the beginning and the end of a sentence makes possible a greater emphasis on meaning, the lack of such a variation may account for the fact that the subjects showing such lack are classed as poor readers.

Continuous Relationship of Eye and Voice.

A comparison of the width of the eye-voice span as measured at different positions in a paragraph showed a considerable variation from point to point in a selection read by a single subject. In order to get more complete data on the continuous relationships of the eye and voice, a very detailed analysis was made of the records of a number of subjects showing the exact relationship of the eye and voice at each word in the selection. Plate II gives a section of a typical record treated by this kind of analysis. The selection has been duplicated in parallel lines in order to show the eye-voice relationship more clearly. The upper line of the pairs may be called the eye line, and shows the position and duration of eye-fixations in the same manner as in the previous plate. The lower line of the pairs may be called the voice line. The diagonal lines connect the positions of the eye and voice for every fixation.

In the record shown in Plate II, the positions of the eye and voice were synchronized first as the voice was pronouncing the word "two" in line 1. As the voice began to pronounce "two" the eye was fixated on the last letter of the word "were," which is the fifth fixation in the line. These two points may therefore be taken as a base of measurement for the determination of the relative positions of the eye and voice at succeeding words and fixations. It will be observed by reference to the plate that 34 (20 + 14) fiftieths of a second elapsed

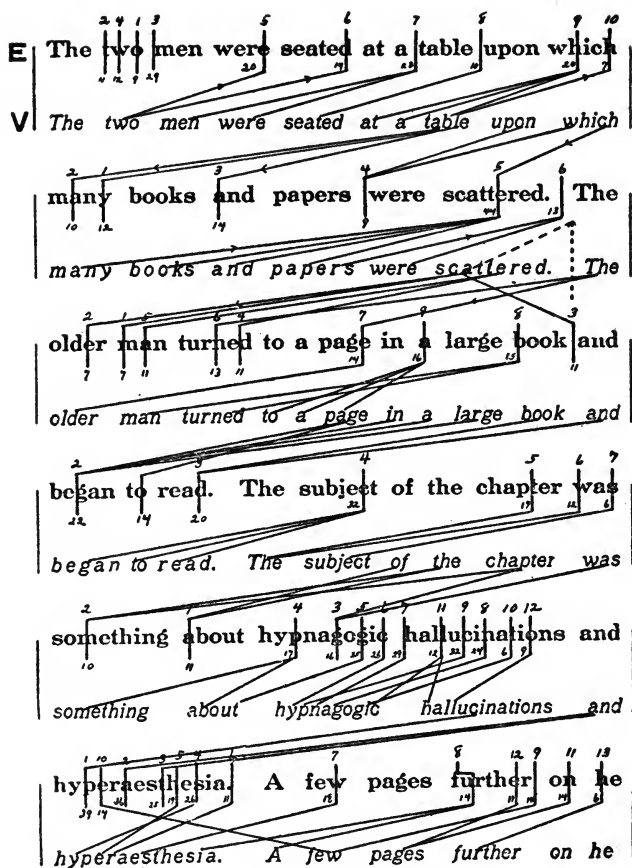


PLATE II.

Showing continuous eye-voice relationship of a good reader from the high school.

during fixations 5 and 6. The time on a stop-watch showed that at a point 34 fiftieths of a second after beginning to pronounce "two," the voice had not begun to pronounce the next word, indicating that fixations 5 and 6 were both made during the time used by the voice in reading the word "two." From fixation 5 to the end of fixation 7 a time of 57 ($20 + 14 + 23$) fiftieths of a second elapsed. The record from the stop-watch and the dictaphone showed that during the interval from 34 fiftieths of a second to 57 fiftieths, the voice had pronounced the words "men" and "were." By a continuation of this

form of analysis the relative positions of the eye and voice were located for every word and every fixation.

Two characteristics of reading are apparent from this type of analysis. The first is the elastic nature of the eye-voice relationship. The width of the spau is evidently varied according to the demands of the content of the material. When the difficult words in the last two lines are encountered there is a very marked modification of the eye-voice span.

A second type of result, which was indirectly found, is an apparent relationship between the length of fixation pauses and difficulties in reading. The average duration of a fixation pause for the subject shown in Plate II is 12 fiftieths of a second. If the three difficult words are again observed it will be seen that most of the fixations upon these words are much longer than the average. This suggested that an explanation of the occurrence of long fixations might be found in the content of the reading selection.

In order to test this possibility a study was made of all fixations which were longer than 20 fiftieths of a second in the records of nineteen subjects. The readings of these nineteen subjects showed a total of 259 fixations which were over 20 fiftieths of a second in length. The selection which these subjects read contained ninety words. If the long fixations simply occurred at random there should be an average of 2.9 for each word in the paragraph. The point of interest is that the long fixations are not distributed in a random fashion with 2.9 falling on each word, but that they occur much more frequently on certain words than on others. The greatest number of long fixations fall upon the words "hyperaesthesia," "hypnagogic," and "hallucinations," these words receiving 31, 22, and 16 fixations respectively.

The interpretation of such results must be related to the development of reading habits. The records of the elementary subjects, whose reading habits are still immature, show that long fixations are frequent and occur all through the selection. Evidently long fixations are characteristic of immature types of reading. The high school and adult subjects represent more mature readers. In general they have outgrown the long fixation habits. However, when words of special difficulty or difficult phrases are encountered they return immediately to the primitive type of habits characteristic of the immature reader.

The Eye-Voice Span and the Recognition of Meaning.

The significant factor about a wide eye-voice span is that it provides a wide unit for the interpretation of meaning before the voice reaction takes place. In order to test this factor a paragraph was constructed containing a number of words which are spelled alike but may be pronounced differently. The section used is shown below, with the test words italicized.

The boys' arrows were nearly gone so they sat down on the grass and stopped hunting. Over at the edge of the woods they saw Henry making a *bow* to a little girl who was coming down the road. She had *tears* in her dress and also *tears* in her eyes. She gave Henry a note which he brought over to the group of young hunters. *Read* to the boys it caused great excitement. After a *minute* but rapid examination of their weapons they ran down the valley. *Does* were standing at the edge of the lake making an excellent target.

The paragraph is so constructed that the meaning of the test words is not evident until the next few words in the sentence are read. For example, in line three, the word "tears" is ambiguous until the word "dress" is reached. The hypothesis in regard to the eye-voice span is that no error will be made in the reading if the span is wide enough to enable the reader to take in the word "dress" before pronouncing the word "tears." If the span is not wide enough to do this there is a strong probability of error. By varying the distance between the test words and the part of the sentence which qualifies it, a rough measure of the width of the eye-voice span can be secured. If this hypothesis is correct, subjects having a wide eye-voice span should make fewer errors than those whose span is narrower. The trial of the paragraph with a group of subjects showed the hypothesis to be correct.

Photographs of the oral reading of the test paragraph brought out two facts clearly. The first is that difficulties in the recognition of meaning are reflected in the eye movements by characteristic types of confusion. This was demonstrated by a comparison of the dictaphone and photographic records. The second fact is that subjects having a wide eye-voice span have less difficulty with such material than subjects with a narrow span.

Photographs were also taken of the silent reading of the test paragraph. The same characteristic eye-confusions occur at some of the test words as occurred in the oral records. Evidently there is in silent reading an eye-recognition span which is similar to the eye-voice span in oral reading. Comparisons of oral and silent records show that whenever difficulties with the test words are experienced, the same types of confused eye movements occur. Since subjects with a wide eye-voice span have fewer difficulties with the oral reading of such material than subjects with a narrower span, it suggests that training for a wide eye-voice span would carry over into silent reading habits.

It has been pointed out that when difficulties are encountered in oral reading the eye-voice span is immediately reduced to a primitive form. The same thing occurs in silent reading. Silent reading records showed that when the difficult words in the test paragraph were encountered, the eye returned to the word causing the difficulty just as it returned to the position of the voice in oral reading. If the difficulty in getting the meaning in silent reading is sufficiently great, there is a reversion not only to the habit of bringing the eye back to the location of the recognition of meaning, but also to the most primitive habit of silently pronouncing the words. This reinstates the most primitive form of reading where the eye, the voice, and the meaning proceed together.

The development of the reading process may be traced through three stages. First, the most primitive or immature stage of oral reading where the eye, the voice, and the meaning are all focused at the same point. Such a stage is illustrated by the reading of one of the second grade subjects where the average eye-voice span was less than the width of the average word in the selection. Second, a more mature stage of oral reading where there is a considerable span between the eye and the voice, with the center of attention near the eye, the voice reaction occurring in a semi-automatic fashion. Third, the stage of silent reading where the reader is entirely relieved of any attention to the voice and where the entire attention can be given to the eye and the meaning, making possible the development of a much higher degree of proficiency.

PROPHECY OF LEARNING PROGRESS BY BETA

GARRY C. MYERS,

Cleveland School of Education, Cleveland (Ohio) Public Schools.

During April of 1920 men of the First Recruit Educational Center, Camp Upton, N. Y., consisting of about 1400 illiterate soldiers, were given Beta. Then these men were reclassified within each grade upon the basis of the Beta ratings. Because of a quarantine in one company and considerable illness in a few other companies a few hundred men were not tested, which meant that the classification was not complete. However, after that time all men entering the Center were tested and placed into classes on the basis of their relative intelligence rating.

For the purpose of administration in that school the course was divided into six grades—not six traditional grades, but six grades for that school. The work covered by each grade as well as the requirements for promotion were highly standardized. Promotions took place every two weeks which was the minimal interval for completing the work of each grade. Because of its method of promotion this school afforded an excellent opportunity to study to what degree Beta ratings prophesy learning progress.

On the basis of all available records of men in the school June 28, 1920, a study was made of the time which had been spent in school by the men as they were found at that time in the respective sections of the several grades. For example, X with a Beta rating of 55 then in grade two, section A, had been in school a total of 31 days (including Saturday and Sunday). The problem was to find out if the low Beta men progress as rapidly as the high Beta men. Of course it were more accurate if the exact time for completing each grade were recorded for each man and if actual school days exclusive of Saturdays, Sundays and holidays, guard days, recruiting-duty days and sick days were first excluded.

Here is what was found:

Median number of days which men of each section of each grade had spent in school up to June 28, 1920.

Grade.	(988 men pretty evenly distributed by sections and grades.)			
	Highest Beta or Section A. Med. Days.	2nd Highest Beta or Section B. Med. Days.	3rd Highest Beta or Section C. Med. Days.	4th Highest Beta or Section D. Med. Days.
1	26	31	37	68
2	37	48	92	110
3	58	110	139	
4	110	139	129	
5	120	137	159	
6	137	155		

The median time in school by those who had completed the six grades at the Recruit Educational Center in May and June including holidays, quarantines, etc., was 173 days.

From the table one reads, for example: "On June 28, 1920, the median time which had been spent in school by Section A (section making the highest Beta rating) was 26 days as against 68 days for the lowest Beta section; the median time which had been spent by the highest Beta section of grade 2 was 37 days as against 110 days for the lowest Beta section. That the median time for the second highest Beta section of grade 4 should be higher than for the lowest Beta section of that grade is probably explained by the fact that the teacher of this section did not conform to the standards set for his section. Instead he introduced considerable experimental subject matter.

Although these figures clearly show that those ratings highest in Beta tend to progress much faster than those rating lowest in Beta, yet the figures do not tell this story well enough. In the first place, since more than fifty per cent of the men were of foreign language, many of whom did not even understand English well on entering school, some of these men relatively bright did not reveal by their Beta ratings their native capacity to learn to the same relative degree that the English speaking illiterate did—for even Beta handicapped the non-English men. This means that some men were classed in a relatively too low section. Likewise the non-English of the highest Beta section, because of English difficulties, made relatively slow progress in the first few grades regardless of their capacity. Moreover, the non-school days, sick days, quarantines, guard days, etc., which were constant for all sections of a grade,

added too many days to the high grade Beta sections in comparison with the low grade Beta sections. If, for example, an average of 10 days were deducted by these demands from the actual time in school for each section of a grade, the ratio of the time in school by the highest to that of the lowest would greatly be increased. Furthermore, the men selected from time to time for several weeks recruiting service naturally drew from the higher Beta-sections, adding thereby an undue proportion to the time-in-school-records, since there were not sufficient data available to make the proper deduction for this time. Finally, it should be added that 68 days for D Section of grade one would have increased materially if the study had been made a few months later unless a large number had been eliminated, for 67 cases of this group on June 28, 1920, had been in school a median of 165 days. The Stanford Binet ratings of these 67 cases were as follows:

Mental age in years.	No. cases.
5—5.9	5
6—6.9	15
7—7.9	24
8—8.9	19
9—9.9	4

Likewise 23 men of the lowest Beta section of the second grade had been in school a median of 217 days. Their mental age ratings were:

Mental age in years.	No. cases.
7—7.9	6
8—8.9	11
9—9.9	6

Of 19 men who had been dropped from school as "hopeless" during May and June the median time in school was 190 days with the following distribution:

Mental age in years.	No. cases.
5—5.9	1
6—6.9	6
7—7.9	7
8—8.9	3
9—9.9	2

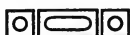
These data seem to justify the action of the War Department in extending the Upton intelligence classification plan to all the Army's Americanization schools.

This plan, whose details are now pretty closely adhered to in all the Recruit Educational Centers, and which after its evolution at Camp Upton, the writer,* under orders, submitted to the War Department, provides that all men of a given Recruit Educational Center be examined by Beta and classified within each grade on the basis of their relative ratings, and that all men thereafter be examined only on entering the Center. It provides that the rating limits for each section be so adjusted that, although the men in the better section shall always be relatively higher in Beta rating than the men of the next lower section, it shall keep the size of the sections practically equal.

No repetition of Beta is made; but in promotion, all men of a given grade are advanced to the appropriate grade regardless of section. Then the names of those remaining in a given section together with those promoted to that section are pooled and ranked on the basis of the original Beta ratings. The number desired for the best section is then counted off from the highest of the group, those for the next section, next, and so on for the entire grade. The clinician and military psychiatrist, of course, takes care of the individual "variables."

*Then Director of Education, First Recruit Educational Center, Camp Upton, N. Y.

DEPARTMENT FOR DISCUSSION OF RESEARCH PROBLEMS



Conducted by LAURA ZIRBES



This department has a two-fold function. It aims to serve research workers as well as educators, whose work brings them in close contact with children in the schools. It hopes to accomplish this service by suggesting research studies, which will meet well-defined school needs.

In order that this service may be real and effective, the co-operation of research workers and school people is desired. Correspondence with reference to the following questions will be considered in selecting topics for future discussions.

- a. Which of the studies proposed would help you to solve a practical problem?
- b. What topics might well be added to this list? Replies may be addressed to: Miss Laura Zirbes, 646 Park Ave., New York City.

AN ADDITIONAL CRITERION FOR THE SELECTION OF THE ELEMENTS OF MENTAL TESTS

J. CROSBY CHAPMAN,
Yale University.

The two criteria which have been chiefly used for investigating the effectiveness of a particular element of a mental test have been

1. The increase in achievement from age group to age group;
2. The variations within each age group (coherence).

The extent to which the above criteria are fulfilled is measured by administering the particular element of the test to large groups with increasing chronological age. The first criterion obviously provides no safeguard against the training factor and the second not as much as might be desired.

In thus establishing the validity of the test elements we have been guilty of loading the dice in our own favor, we have made our task too easy. The factors which contribute to an increase in the achievement of the higher age group are—

1. Increase of intelligence;
2. Longer period of exposure to environmental and training influences. However important the second factor may be as an indication of effectiveness, it is a disturbing factor in determining the reliability of the test element as a measure of pure intelligence, defined

as the power to adapt to a novel situation. It is essential that we set up an additional criterion which will load the dice in the opposite direction. Whereas in the above situation both factors work in the same direction, we must set up a criterion in which they work in opposite directions.

That such a criterion is possible the following simplified case will illustrate. Suppose instead of merely trying out the test element on large groups of increasing mental age, in addition each element of the test was submitted to a group of chronologically young, but mentally bright children, and to a second group of children, mentally dull but chronologically old. To make the situation more specific, let us suppose that each test element is tried out on

1. Fifty children of age 8 who show on any composite test a mental age of 10.

2. Fifty children of age 12 who show on any composite test a mental age of 10.

We should then establish as our additional criterion of the excellence of a particular test element the extent to which the younger group exceeded the older group.* In this case the environmental factor would be working against the intelligence factor. In other words we would have loaded the dice against ourselves. Could we not in this manner arrange our tests in ascending order of merit, according to the way in which they differentiated more and more successfully in favor of the younger group? Would it not happen that the boasted merits of the arithmetic tests and vocabulary tests would very much dwindle, possibly melt into thin air, in favor of those tests which less obviously are subject to environmental factors? In this way we might conceivably eliminate certain tests which measured by present criteria give good indications of differentiating power but which under more refined testing would show their limitations.

A more rapid method of attacking the problem would be to compare the scores obtained in any of the recognized composite group tests by the young-bright, and the old-dull children. We could take the chronological ages as far as possible apart, consistent with securing a reasonable number of each of the groups scoring approximately the same number of points. We could then investigate the extent

*See the work of Chatzen on a less exacting criterion.

to which these two different groups had secured their totals on the same elements of the test.

The author is at present collecting and reviewing experimental data which will shortly be published with a view to testing the extent to which the application of this "criterion of environmental influence" will change the present estimation of the relative effectiveness of the various test elements. The publication of this note has been hastened by the consideration that many individuals have the necessary data, in greater quantity than the author, with which to test the fruitfulness of this exacting criterion.

Unless the above criterion is applied we get the anomalous situation which at present exists, that the mental ages of children in a school system are a function of the enlightenment of that school system with reference to retardation and acceleration! It is obviously absurd to prejudice the mental rating of an extremely bright child simply because a cast iron school system has, through ignorance, denied the promotions necessary for a certain facility in, let us say, the arithmetic and vocabulary tests which contribute to high mental standing! The eventual outcome of the changes in the mental test elements must be to prove that the brighter children are really more intelligent than the present tests indicate, and that the duller children are even less capable than present measurements reveal.

HOW CAN GROUP TESTS BE BETTER ADAPTED TO VERY YOUNG CHILDREN?

Dr. Agnes Rogers remarked in a recent lecture at Teachers College that some makers of group tests for children of kindergarten age had made some very ingenious provisions for securing attention and enlisting interests. She also believes that the consideration of the following proposals in determining the content or form of such tests might add materially to the successful administration of group tests to young children:

1. That when pictures are used, their size and clearness be determined psychologically.
2. That the content take more careful account of the child's range of experience.
3. That the tests be divided into parts, printed on separate small sheets or cards, to

avoid the confusion attending the presentation of a large number of pictures or an unwieldy leaflet. 4. That there be a standardized requirement for giving separate parts of the tests at intervals rather than in immediate succession. 5. That there be a standardized period for practice in the use of paper and pencil as a test preliminary for children of kindergarten age especially. 6. That there be some method to reduce the tendency to communicate, which is especially difficult to control in absurdities tests.

Dr. Rogers thinks that the group test is particularly adapted to the measurement of the timid child, especially when individual tests are administered by strangers.

The present writer suggests that there be some group activity immediately preceding the test, so that pupils unaccustomed to group responses are not unduly hampered by the requirements of group testing. This could be in the nature of a prescribed game in which pupils follow instructions without conversation or individual variation. Such group activity would be part of the standardized test instructions and lead to more uniform responses. The objective evaluation of the curriculum and methods used with young children will no doubt be facilitated by the use of tests which take careful account of the limitations of pupil material.

L. Z.

Dr. Ernest Horn of the University of Iowa and Dr. William McCall of Teachers College, Columbia, will contribute discussions to this department in the next issue.

NEW PUBLICATIONS IN EDUCATIONAL PSYCHOLOGY AND RELATED FIELDS OF EDUCATION

1. *A new monograph on the psychology of reading.* Dr. Buswell's study¹ of the eye-voice span is the fifth of The University of Chicago monographs which deal exclusively with reading. The laboratory results which it reports throw much light on problems suggested in studies by Quantz, and Judd, and C. T. Gray. With the apparatus used by Gray the author obtained photographic records of the eye-movements of fifty-four subjects, selected to include equal numbers of good and poor readers, from the second grade through first year college. By the simultaneous use of the camera, dictaphone, and ingenious synchronizing devices, the records were made to reveal the eye-voice span, or the point at which the eye was fixated when the voice pronounced any given word in the selection.

The study was very carefully set up to reveal any significant relationships which might exist between eye-voice span and other elements of reading ability. Plates, figures, and tables are presented so that the data may be consulted by the reader and conclusions can be studied by reference to the data on which they are based. There is cumulative evidence that the element studied has significant relations to comprehension, rate, and the quality of reading, as well as to number and length of fixations. The data lead to the conclusion that eye-voice span varies for the same reader within a given selection for causes which have to do with content and meaning, rather than with position in the line. Mature readers have longer spans, which give them a distinct advantage in the recognition of meaning and the avoidance of error, as demonstrated by tests containing words difficult to understand or interpret. When difficulties cannot readily be negotiated, mature readers revert to the primitive type of habits characteristic of the immature reader.

The study shows that the eye behaves similarly when encountering difficulties in silent reading. The immature reader evidently derives

¹Buswell, G. T. *An Experimental Study of the Eye-Voice Span in Reading.* Supplementary Educational Monographs, Vol. III, No. 17, Chicago. University of Chicago, Dept. of Education, 1920. Pp. XI + 105. \$1.00. Paper.

less benefit from silent reading as his eye and voice proceed more nearly together. There is hence no opportunity to adjust between seeing and saying. If this is a measure of immaturity, Dr. Buswell's study leads to the conclusion that the advantage of a wide attention span should be made available by special training during the early grades. The laboratory studies show that good readers in the lower grades have the eye-habits of mature readers. The extent to which a wide span may be developed, or to which it is dependent on native capacity, is a problem well worth careful investigation.

L. Z.

2. *A New Handbook on Sex Education.* While Dr. Galloway's new book² is intended primarily for the use of college men there is much in it which should give it an audience in other circles. The author feels that the college man will be much more likely to develop a satisfying point of view in matters pertaining to sex if he considers his responsibility to younger boys and prepares himself for leadership in the character development of others. The organization of the book is noteworthy. Discussion topics are listed at the beginning of each chapter. An introductory statement is then followed by brief specific questions and answers, with references to source materials.

The introductory statements taken together are a sane and clear statement of the issues, replete with well chosen illustrations which cannot fail to impress the careful reader with the fundamental philosophy upon which the author builds his sex-education program. The dynamic character of appetites and desires must be considered in any thoroughgoing plan for character development. Our reactions to appetites and impulses are measures of our control and character, and are a means of self-development which cannot be side-stepped or eliminated without grave consequences.

Only ignorance of the social and ethical potentiality of refined and redirected appetites and desires leads to a social and educational program which capitalizes the dynamic human factors underlying behavior. Dr. Galloway's book makes its greatest contribution by clarifying the issue, and showing its relationship to individual and social progress.

L. Z.

²Galloway, T. W. *The Sex Factor in Human Life.* The American Social Hygiene Association, New York City, 1921. Pp. 142.

3. *A report of the use of mental tests in school administration.* The aim of this study^a is stated very simply: to show how the results of certain educational and psychological tests may be used as aids in problems of school administration. The results reported were obtained from tests of 125 boys and girls of the Junior Division of the University High School, Eugene, Oregon. These pupils were drawn from two types of homes:—some came from families who were professionally connected with the University; another group was drawn from the homes of unskilled workmen.

All the pupils in the school were tested with the Stanford Revision of the Binet Scale, the Army Group Examination Alpha, and the Chicago Group Intelligence Test, devised by Freeman and Rugg. The results for the total group, stated in terms of the median, are as follows: Mental Age, 14.9; Intelligence Quotient, 107.0; Alpha, 89.0; Chicago, 38.0. These results show clearly that the children of this group are a selected group, with intelligence above that of the average for this age. When the results of each group test are correlated with the Binet scores, the correlation for Alpha is found to be 0.73, in comparison with 0.62 for the Chicago Scale. The author thinks this difference may be accounted for by the fact that the Chicago scale requires 15 minutes less time than does the Alpha test.

Six experienced teachers, after having received some instructions in methods of making intelligence ratings, rated all the pupils they each knew. These estimates of intelligence were correlated with the I.Q.'s. The resulting coefficients ranged from 0.61 to 0.72, with an average of 0.68. The pooled ratings also gave an r of 0.68. This high correlation is contrasted with a lower one of 0.50, which results from the correlation of the estimates of ten inexperienced practice-teachers, with the I.Q.'s.

The question of acceleration and retardation is discussed, and it is shown that, on the basis of Mental Age, 31.1 per cent are accelerated, 22.7 per cent are normal and 46.2 per cent are retarded. Of 36 pupils who are retarded by usual (actual age) standards, 41.7 per cent are accelerated, by Mental Age. Of 40 who are accelerated by the usual standards, 55.0 per cent are retarded, by Mental Age. The author comes to the conclusion that "the accelerated pupils upon the basis of actual age tend to become retarded upon the basis of

^aRuch, Giles Murrel. *A Study of the Mental, Pedagogical and Physical Development of the Pupils of the Junior Division of the University High School, Eugene, Oregon.* University of Oregon, Eugene, Oregon.

mental age, and that the reverse holds for the group considered retarded by the usual standards"—a conclusion which other studies corroborate.

Seven tests of educational achievement were applied to the group:—Courtis Standard Research Tests in Arithmetic, Series B; Stone Reasoning Tests; Gregory Language; Ayres Spelling; Handwriting scored by the Ayres Scale; Kansas Silent Reading; Douglas Tests for elementary Algebra. The results for grades VII and VIII are compared with standard scores; they prove to be "well above the norms in all tests except in the fundamentals of arithmetic as shown by the Courtis Tests, and in writing." A study of the Stone Reasoning Tests seems to indicate that the low scores in computation are due to poor training in the lower grades of the system. The poor showing in handwriting is explained by the fact that too little time on the school program is allotted to penmanship practice.

Certain anthropometric measurements were obtained to supplement the above results. In height and weight no significant deviations from the available norms were found. When compared with Smedley's results the Eugene boys and girls show "markedly greater development of lung capacity at all ages." From certain correlations obtained, the author concludes that there "exists a positive relationship between the vital index and mental ability. This relationship is not very perfect and is probably of indirect value."

The most important of the conclusions based upon the results just discussed are as follows:

1. While it is true that trained teachers of long experience can estimate intelligence with a high degree of accuracy, they do occasionally make serious errors. It is in these exceptional cases that intelligence tests are most needed. "The fact that comparisons of the ratings obtained by the two methods show discrepancies leads to deeper analysis of specific abilities by teachers, and to a truer understanding of the complex relationships between the many factors combining to determine school achievement."

2. The results of such tests may be of great assistance in helping to determine the speed of school advancement of each individual. If forced promotions are necessary, the results of the tests will aid in selecting the pupils for such promotion. Especial provision, in separate classes, should be made for children of high native ability.

3. Such tests are of value in determining the *special* needs of each child.

4. "Mental deficiency is a far more powerful factor in the cause of retardation than has been commonly supposed."

5. "Correlation of abilities in school subjects with general intelligence shows the highest coefficients in the language and reasoning tests, and the lowest in spelling and writing. School marks and arithmetical ability are intermediate."

6. Pupils who are to be classified as of average ability are "somewhat more likely to realize results in school according to their ability than are either very able or very dull pupils."

7. "In view of the value of the results of intelligence testing from the administrative angle the amount of time required for tests cannot reasonably be held to be prohibitive."

G. L. COY.

Ohio State University.

4. *A survey of material used for memory work in city public schools.* This bulletin attempts to present the present day American city public school practice in regard to the specific material used for memory work and the amount of such work required. There are given in detail the method of procedure and the results of a survey of fifty city courses of study, selected from a much larger group, primarily on the basis of complete lists of memory material offered and the specificness of the memory work requirements. Due to indefiniteness in the courses of study which were surveyed, only the data which refer to poetry memory work are at all complete.

From the complete lists of 2,435 poems mentioned in the various courses of study, the investigator selects the 329 that were mentioned in five or more courses. These selected poems are presented in various suggestive and useful tabular forms. The first of these tables gives the titles and the poems arranged in the order of frequency of mention and shows, as well, the number of different cities mentioning, and a weighted value for each poem. This value is obtained by using the factors 1, 2, 3, and 4 respectively for mentions in which the poem is given "for study"; is suggested for memory, other poems being required for memory; "suggested for memory without distinction"; "required for memory." A second table shows

⁴Bamesberger, Velda C. *Standard Requirements for Memorizing Literary Material.* University of Illinois, Bureau of Educational Research. Bulletin No. 3. Pp. 93.

the distribution of the mentions of each poem according to the school grade for which it is recommended by the various courses of study. The results of the final compilations of the data in this form are then so presented as to show just which poems are now tending to be used for memory work in each grade from the first to the eighth inclusive. A group of "Preferred List of Poems for Memory Work" arranged by grades and consisting of 112 poems presents, we are inclined to agree with author, "the best short lists which an investigation of present courses of study can yield." These preferred lists represent in the main the agreements in graded classifications of poems as ascertained from the data of this investigation and those of a similar investigation made by Lewis Atherton and published in 1914.

These lists should prove suggestive and helpful to all teachers. To those teachers who have not sufficient literary training, or initiative, or time left them by the frequently overburdensome schedules, these lists may well serve as excellent guides to keep them within the limits set by the accepted standards for this type of work. Those teachers who wish to make use of the lists, as well as the educationalists who may desire to do experimental work upon the relative values of the various poems listed, will discover that much aid will be given them by the excellent arrangement of the poems into a "Finding List" arranging the poems by author and giving titles, first lines, suggested grade, and text references.

We feel very strongly, however, that every teacher making such use of these lists should be at the same time alive to the inadequacies inherent in lists thus compiled. The author is, herself, well aware of the practical limitations of the lists. She emphasizes the fact that the lists are "based on agreement between courses of study. If there is any constant bias—any convention which has grown up whereby one writer of a course imitates others—these results will likewise be biased in the same direction. If, for example, there is a tendency to assign to elementary school children poems which they cannot understand or appreciate, this tendency will be evident in the resultant lists."

Certain tendencies, which might well be considered as biased tendencies, are clearly shown by the author to influence the composition of the lists. There is the tendency to unduly favor American at the expense of English authors, shown not only by the many more

poems chosen from American authors, but also by the inclusion of work from a large number of minor American poets. The work of foreign language writers is practically excluded, as is also the work of recent poets.

In the editorial introduction is found the suggestion "that any desirable form must be based upon present practice as the point of departure." Perhaps, after all, the chief value of this present work may be found to consist in supplying this necessary point of departure.

HARRY W. CRANE.

Ohio State University.

5. *An Experimental Study of Memory.* Using a variety of materials (words, geometrical forms, proverbs, and syllables), with approximately 100 adults, and 600 school children, as subjects, Dr. Achilles has made an intensity study of Recall and Recognition.^a As in earlier investigations, it is found that the number of items recognized surpasses the number recalled by amounts depending upon the material and subjects. With relatively homogeneous groups, a low but positive correlation between the two processes is found. Other results conform interestingly to recent theories that mental behavior depends upon many, more or less specific, rather than a few general capacities. These correlations between Recall for different types of material are low (averaging around 0.10 with large P. E.'s). The correlations for Recognition of different materials are similarly low but positive. Women and girls are in general found to be slightly superior to men and boys. Both Recall and Recognition increase rather uniformly with age (8.5-16.5 years) and with grades (4-8). The younger pupils in a grade usually surpass the older. The tests were applied to insane patients but produced results of no diagnostic significance. A final chapter gives the details of the Recognition process, showing in general that the subject is more correct in judging a thing *as not seen* before than *as seen* before.

A. I. G.

^aAchilles, Edith Mulhall. *Experimental Studies in Recall and Recognition*. Archives of Psychology, 1920. No. 44, pp. V+80.

THE JOURNAL OF EDUCATIONAL PSYCHOLOGY

Vol. XII

MAY, 1921

No. 5

A SURVEY OF THE THREE FIRST GRADES OF THE HORACE MANN SCHOOL BY MEANS OF PSYCHOLOGICAL TESTS AND TEACHERS' ESTIMATES, AND A STATISTICAL EVALUATION OF THE METHODS EMPLOYED.

CLARA F. CHASSELL,

Psychologist of the Horace Mann School, Teachers College, Columbia University,

and

LAURA M. CHASSELL,

Instructor in Psychology, Ohio State University.

Part I of this article, published in the February issue of this magazine, contained a statement as to the causes which led to the undertaking of a survey of the three first grades of the Horace Mann School, and presented in full the data obtained in that survey. The measures employed included psychological tests, both group and individual, and rankings of the pupils by the teachers in maturity and ability in reading. In addition, it suggested a plan for utilizing the data thus secured, for the purpose of reclassifying the pupils into relatively homogeneous groups.

Part II continues the report of this survey, recording the correlations obtained between the various measures, evaluating these measures by comparing them with a composite of all the measures utilized, and giving a detailed account of the statistical methods employed in the conversion of these measures into mental ages.¹

¹For guidance in statistical method and evaluation of measures reported in Part II, the writers are indebted to Prof. T. L. Kelley, formerly of Teachers College, now of Leland Stanford University, and Prof. H. A. Ruger of Teachers College.

PART II. CORRELATIONS, EVALUATION OF MEASURES, AND STATISTICAL PROCEDURE

Presentation of correlations and evaluation of measures.—Table V presents the correlations obtained² between the various measures in the survey, including the composite of all. It should be noted that with the exception of the reliability coefficients reported³ these are based on mental age measures only.

TABLE V
Correlations between Mental Ages Secured by the Various Measures
Employed in the Survey

	Stanford Revision	Pressey Primer Scale	Meyer Tests	Com- bined Group Tests	Teachers' Esti- mates	Composite
Stanford Revision.....	(.93)	.45	.53	.64	.72	.89
Pressey Primer Scale.....	.45	.52 ⁴	.37			
Meyer Tests.....	.53	.37	.59 ⁵			
Combined Group Tests.....	.64				.42	.87
Teachers' Estimates.....	.72			.42	(.61)	.85
Stanford Revision and Group Tests96
Stanford Revision and Teach- ers' Estimates.....						.91
Group Tests and Teachers' Estimates77					.98
Composite89			.87	.85	

²According to the product-moment method of Pearson.

³The reliability coefficients actually taken into account in determining the weightings assigned to the various measures entering into the composite, referred to in Part I, were only approximate in the case of the Stanford Revision and the teachers' estimates. The coefficients reported in Table V were variously ascertained. The figure given for the Stanford Revision (.93) is the correlation found between earlier and later tests. (See Terman, L. M. *The Intelligence of School Children*, p. 142.) The figures for the two group tests (.52 for the Pressey and .59 for the Meyer) were calculated by computing the correlation between the sum of the scores for odd tests and the sum of the scores for even tests. These coefficients are raised to .68 for the Pressey and .74 for the Meyer by the application of 'Brown's' formula. (See Brown, William, *The Essentials of*

Mental Measurement, p. 102, fn. The formula as there stated is $r_n = \frac{1 + (n-1) r_1}{n}$

In this instance we are not interested in determining the number of applications of the test necessary to give any desired degree of reliability, however, but in determining the reliability of the entire test from the reliability of one-half of the test. Thus, in the case of the Pressey the correlation of .52 between the sum of the scores for odd tests and the sum of the scores for even tests is the reliability coefficient for only one-half of the test. In determining the reliability coefficient of the entire test, $n=2$. Applying the

formula, we have r_n , i. e., the reliability coefficient of the entire test, $= \frac{1 + (2-1) .52}{2} = .68$.

Vol. III, p. 281. The correlation given for teachers' estimates (.61) is that found to obtain between estimates of school work, made by two or more teachers of the same children in the Horace Mann School, at the time the criterion for the National Research Council Tests was being compiled.

⁴Raised to .68 by the application of 'Brown's' formula.

⁵Raised to .74 by the application of 'Brown's' formula.

An examination of the correlations reported for the Stanford Revision shows that the highest correlation between this measure and any other single measure appears in the case of teachers' estimates, namely, .72 (60 cases being included).⁶ In this connection the findings of two other experimenters, while not directly comparable with our results,⁷ are sufficiently related to be of interest. Terman⁸ reports a correlation of .48 between intelligence quotients and teachers' estimates of intelligence, the rankings being on a scale of 5. Similarly, Dickson⁹ found a correlation of .79 between these same measures in the case of 149 first grade children, the teachers having been expressly cautioned to take the children's ages into account in making the ranking.

The correlations with the two group tests¹⁰ are much lower, namely, .45 for the Pressey Primer Scale, and .53 for the Meyer Tests¹¹ (based on 37 and 45 cases, respectively). A correlation of only .37 was found between the two group tests themselves. A combination of these two tests into a single measure results in a correlation of .64 with the Stanford Revision.

The correlation of .42 between the combined group tests and the teachers' estimates is surprisingly low, since a group examination, to a far greater extent than an individual, would seem to require a response similar to that met by the teachers in the usual classroom situation.

From the practical standpoint, the correlation of greatest interest is probably the one between the Stanford Revision and group tests combined with teachers' estimates. The correlation of .77 between these measures, based on only 30 cases, is sufficiently high to raise the question as to whether an individual psychological examination is really necessary for determining classification and promotion. If the possibility of partially explaining this relatively high correla-

⁶It should be borne in mind, however, that the teachers were already familiar with the intelligence quotients of a large number of their children. While it is probable that no direct comparison with these quotients was made, this previous knowledge doubtless had some influence upon the rankings assigned.

The rankings made by the teachers in our survey were of maturity and ability in reading, and in general extended over the entire range of the class.

⁷The Measurement of Intelligence, p. 75.

⁸See Terman, L. M., *The Intelligence of School Children*, pp. 51-52.

⁹All correlations reported with the combined group tests are based on 30 cases only. This small number of cases in which direct comparison is possible is due to the prolonged absence of an unusually large number of children.

¹⁰As explained in Part I, the Meyer Tests used in this survey are those constructed by Miss Helen Meyer, described in an unpublished thesis, "Group Tests for Grades I and II," on file in the Columbia University Library.

tion by the fact that the teachers were familiar with the intelligence quotients of a large number of their children be entirely overlooked, it should still be emphasized that a correlation of .77 means a reduction in the ratio of the variability around the regression line to the variability around the average of only .64.¹² Even if this reduction in variability were much greater, in view of the great value of the Stanford Revision for the understanding of individual pupils, we should still hesitate to say that it was not an essential instrument for satisfactory classification and promotion. Certainly, however high this correlation may be found by other investigators, the Stanford Revision or its equivalent will continue to be indispensable at least for determining the placing of the problematical cases of any kind.

Aside from such information as may be afforded by the correlations with the Stanford Revision, already reported, for an evaluation of the various measures used in the survey we are dependent upon a comparison of these measures with the composite of all.¹³ As explained in Part I, this composite is the average mental age found by adding the mental ages obtained from the Stanford Revision, the two group tests, and the teachers' estimates, the Terman mental age being doubled, and dividing this total mental age by the number of measures available for each child, the Terman mental age being counted as two measures.

As judged from the correlations with this composite, presented in Table V, any one of the three types of measures which were used in the survey, that is, an individual examination or two group tests or two teachers' estimates, would approximate the same results for classification and promotion as the composite itself. The correlations for these three measures, based on 60, 30, and 60 cases, respectively, are .89, .87, and .85. In interpreting these correlations, as well as the other correlations to be reported with the composite, it must be borne in mind that the measures compared are not independent, but are already contained within the composite itself.

¹²Obtained by the formula, $\sqrt{1-r^2}$.

¹³It is impossible to evaluate this composite by comparing it with an altogether independent criterion of the value of the various measures employed for purposes of classification and promotion. Such an independent measure would be available subsequently if the results of instruction in groups classified on the basis of the composite could be compared with results obtained in control groups not so classified. One evidence that the composite employed is a fairly satisfactory criterion by means of which to evaluate the measures used in the survey, however, is the general satisfaction felt by the teachers with the results of the survey as summarized in this composite.

Hence they are far higher than would otherwise be the case." Even so, taking these correlations at their face value, since the ratio of the variability around the regression line to the variability around the average would be reduced approximately only one-half¹⁴ by the use of any one of the three types of measures, taken alone, they can be claimed, if used singly, to possess only a limited value for purposes of classification and promotion as compared with the composite.

The correlations with the composite are naturally very much higher still when any two types of measures are combined for purposes of comparison with it, on account of the fact that the four elements included in the two types of measures (e.g., the Terman mental age taken twice plus the mental ages obtained from the two group tests) are identical with four out of the six elements in the composite. Thus the Stanford Revision combined with the group tests gives a correlation of .96 with the composite, the Stanford Revision combined with the teachers' estimates, of .91, and the group tests combined with the teachers' estimates, of .98,¹⁵ (30 cases being included in each instance). The ratio of the variability around the regression line to the variability around the average for these correlations is, respectively, .28, .41, and .20. In interpreting these results, it should be noted that the presence of common elements in the measures correlated, which served in the first place to increase the correlations with the composite, results now in correspondingly lower regression values. Should the special conditions under which these correlations were obtained remain, if only two types of measures are to be used for purposes of classification and promotion, it would seem from these results to make little difference whether the Stanford Revision combined with group tests or group tests combined with teachers' estimates were selected.

¹⁴The reader who is interested in following up the implications of this statement statistically, may be referred to Yule, G. U., *An Introduction to the Theory of Statistics*, ch. XI. (See especially Exercises 6 and 7, p. 227.)

¹⁵Exactly one-half (i. e., 50) if the correlations were all .866. The exact regression values obtained by the formula, $\sqrt{1-r^2}$ for these three correlations are .46, .49 and .53, respectively. It is apparent from this formula, however, that since these regression values involve the correlations with the composite, they are too low for the same reason that the correlations upon which they are based are too high.

¹⁶This last-mentioned correlation may have been artificially increased by still another factor, since as already stated, the teachers' judgments probably were influenced to a certain extent by previous knowledge of the intelligence quotients of the children. The extent of such influence can not be determined. Whatever it may have been, however, it did not succeed in raising the correlation between the Stanford Revision and the teachers' estimates above .72.

The statistical procedure utilized in the conversion of the measures into mental ages.—Reference has already been made in Part I¹⁷ to the fact that the incorporation of such varied data as were secured in the survey necessitated the reduction of all measures to a common basis. The one selected for this purpose was that of mental age.

Before concluding the report of the survey it is thus necessary to present in detail the statistical procedure involved in converting the various measures used into mental ages. The methods used are given in order for the measures employed.

1. The Stanford Revision.

In the case of the Stanford Revision the only computation necessary was that required on account of the fact that the giving of the individual examinations had extended over a long period. It was thus necessary to determine the mental ages of all the children at some specified time. Since the group tests had been given during that same month, the date selected was January 1. After the chronological ages of the children on that date had been computed, the corresponding mental ages were readily obtained by using the intelligence quotients already determined as multipliers.¹⁸ This process was facilitated by the use of an I. Q. slide rule.¹⁹

2. The Pressey Primer Scale and the Meyer Tests.

The method used in converting the scores made in both the Pressey Primer Scale and the Meyer Tests into mental ages, consisted of replacing the score made by a given child in one of these tests by the corresponding mental age, as shown in a table constructed with the median scores in the test for the various ages as a basis.

Since only year norms were available for the Pressey tests and no norms whatever for the Meyer, before the conversion into mental ages was possible it was necessary to build up a table of norms by months. In the case of the Pressey the procedure was as follows: First, since the various age norms provided for the Primer Scale actually represent the typical performances of children six months in advance of the ages specified, the norm for any given year being

¹⁷See *Journal of Educational Psychology*, Vol. XII, No. 2 (Feb., 1921), p. 74.

¹⁸This method is based on the assumption of the constancy of the I. Q., which, although not thoroughly established, seemed to be sufficiently so for use in the present instance.

¹⁹Published by the Reed College Co-Operative Store, Portland, Oregon.

based upon the scores made by all the children who have passed the birthday indicated and have not yet reached the succeeding birthday, the median score given as the norm for six years of age, for example, was taken to have a value in terms of mental age of six and one-half years. The amount of the interval between each of the median scores for the succeeding ages was then determined, and this amount divided by twelve in order to obtain the increment in terms of score which might be taken to correspond to a month of mental age. Each increment was then successively added to the appropriate year norm. The values thus secured were subsequently entered in a table opposite the equivalent year and month of mental age.

The procedure followed in the case of the Meyer Tests was naturally more complicated. No norms being available, it was necessary to estimate norms by means of a comparison of scores made by children in these tests with scores made by the same children in some other test. For this purpose the records from the Stanford Revision were utilized.

First, the mental ages of the twenty-nine children given the six Meyer tests in the regular manner,²⁰ all of whom had also been given the Stanford Revision, were arranged according to the respective chronological ages. Then for each successive half year of chronological age the median of the chronological ages and of the mental ages was found. Each median mental age was then divided by the median chronological age corresponding, and the median intelligence quotient for the children of each half year of chronological age found. The scores made by these same children in the Meyer Tests were similarly tabulated according to the chronological ages of the children, and the median score for each half year ascertained. The age medians thus secured were next divided by the intelligence quotients obtained as described above, in order to secure medians which would be typical of children in general, and which thus might serve as norms for the corresponding chronological ages. The values resulting were then employed in the same way in which the Pressey age norms had been utilized, and a table of scores with their

²⁰Sometime before plans had been made for the survey, two out of the six tests in the Meyer series had been given to the children in one of the rooms. Although, after the remaining four tests had been given to these children, their scores were adapted to make them comparable with those made by the other children, these records were not utilized in estimating the norms.

accompanying equivalents in terms of years and months of mental age built up.

3. The teachers' estimates.

The conversion of the teachers' estimates into mental ages was made by adapting two different methods. Only the general outlines of the methods will be indicated here; the adaptations employed were made in the course of the practical application of the procedures, and would not be of general interest. Both methods involved the transmutation of a given teacher's rankings in each trait, into an equivalent mental age, determined on the basis of the Terman mental ages on January 1st, already calculated, on the principle that for any group of children the distribution of mental age indicated by any measure would be similar to that already found by the application of the Stanford Revision. The first method, which is based on the assumption that the form of distribution concerned is that of the normal probability surface, was employed in the case of the two classes for which all or practically all of the Stanford Revision records were available; the second method, which makes no such assumption, was employed in the case of the third class because the data were relatively incomplete on account of the prolonged absence of a number of the children. For purposes of comparison a third method, used in the survey made in the spring throughout the elementary school, in which the ranks assigned were treated as gross scores, is included.

Method A, in which the distribution was assumed to follow the normal curve. The average and the standard deviation from the average of the Terman mental ages for the class concerned were first computed. Then the number of children assigned each of the teacher's rankings²¹ was determined, and the percentage that this number represented of the total number of the children in the group computed. The average distance of each percentage from the central tendency of the total group in terms of multiples of the standard deviation was then ascertained by reference to Table 54 in Thorndike's *Mental and Social Measurements*.²² The multiples thus obtained were multiplied by the value of the standard deviation from

²¹In a number of instances more than one child had been assigned the same rank.

²²See p. 221 ff.

the average of the Terman mental ages already obtained. The resulting quantities were then added algebraically to the average Terman mental age for the class, and the resulting figures replaced in each case by the nearest whole number. This figure, which represented the number of months of mental age desired, was then converted into years and months.

Method B, in which the distribution was treated as rectangular. In the first place, the highest rank was replaced by the highest mental age found for the children concerned, as its equivalent; and, similarly, the lowest rank was replaced by the lowest mental age. The range in months between the highest and the lowest age was then divided by $(n-1)$ the number of ranks, in order to determine the number of months of mental age which should be taken as equivalent to the interval between two successive ranks. The amount thus secured was subtracted from the highest mental age to secure the value of the second highest rank; then subtracted from this value to secure the value of the third highest rank, and so on, until a value had been found for each rank. The quantities resulting were in each case replaced by the nearest whole numbers, and these numbers, which represented the number of months of mental age appropriate to each rank, were then reduced to years and months.

Method C, in which ranks assigned were treated as gross scores. In the first place, since records for identical children in both measures, only, were usable, records for any children who had not been included in the teachers' rankings and examined by the Stanford Revision as well, were eliminated from the calculations. Next, the remaining ranks being treated as gross scores, the average and the standard deviation from the average of these ranks were computed. Similarly, the average and the standard deviation from the average of the Terman mental ages for the same children were found as in the first method described above. Then, the standard deviation of the ranks being considered as equivalent to the standard deviation of the mental ages, the number of months of mental age equivalent to the interval between two succeeding ranks was found. Thus, if the standard deviation of the ranks of a given class was 3, and the standard deviation of the mental ages for the same children was 6 months, the interval between two succeeding ranks would be considered equivalent to two months of mental age. Finally, with the

average mental age as the starting point, the appropriate number of months was added or subtracted for each following or preceding rank; the resulting figures, which represented mental ages, were replaced by the nearest whole numbers, and the latter reduced to years and months.

Robert S. Ford

THE SCIENTIFIC EVIDENCE ON THE HANDWRITING MOVEMENT

FRANK N. FREEMAN,
The University of Chicago.

THE ARM MOVEMENT DOGMA.

The teaching of handwriting is dominated by a very widespread dogma concerning the best way to write and the best way to teach writing. This dogma is the formulation of the opinion that the arm movement, or the so-called muscular movement, is a superior method of writing, and that writing should be taught by emphasizing this arm movement by giving exercises which develop it and fixing the child's attention upon it. Few writing supervisors question the correctness of this dogma. School administrators, in general, accept the dictum of the supervisors. Some teachers, however, have come to question it, as a result of their firsthand experience in trying to teach this movement. Scientific evidence refutes it almost completely. Even casual observation, if it were made without bias, would seriously undermine it.

The strongest refutation of the arm movement dogma is the fact that with all the efforts which are made to teach it, the pupils do not, in the main, acquire it. It is almost a flat failure. An investigation¹ of 273 children, most of whom had been given strenuous drill in arm movement writing, brought to light the proportion of those who acquired the movement with enough thoroughness to be used by them in their ordinary writing. The analysis of the writing movement was made by means of an instrument which gave a tracing of the movement of the arm. The amount of arm movement could be estimated from this tracing according to the degree to which it corresponded to the original writing. The degree of correspondence was estimated by comparison with a scale of five specimens. The first specimen of this scale is designated by zero and represents no arm movement whatever. The specimens from 1 to 5 represent varying degree of arm movement, 5 being the highest. In specimen 3 the

¹The data referred to in this article are presented in detail in the author's monograph, "The Handwriting Movement," Chicago, 1918, The University of Chicago Press.

arm movement is sufficiently pronounced to cause the hand tracing to resemble somewhat the writing, but not sufficient to enable it to be read.

The results of this study have been tabulated by ages. We may take the data for ages 8 and 14 as representative. These are shown in Table I.

TABLE I.

Age.	No.	Grade of Arm Movement.					
		0	1	2	3	4	5
8	33	.30	.58	.12			
14	38	.03	.42	.26	.11	.11	.08

It will be seen that practically none of the eight-year children used more than the very most rudimentary type of arm movement. Grade 1 of the scale represents simply a slight upward and downward oscillation of the arm. Even Grade 2 represents a very slight amount. These figures indicate that it is almost out of the question to expect to develop the arm movement in primary grade children.

The showing of the fourteen year old children is not much better. Seventy-one per cent exhibited less arm movement than is represented by Grade 3. Practically none exhibited complete arm movement.

The gist of the matter may be put in more concrete form by a quotation from one of the eighth grade pupils who was a subject in the experiment. This pupil was asked to write under a motion picture camera. In order that she might know what was wanted, she asked, "Do you want my ordinary writing or my arm movement writing?" The pupil did not use the term "arm movement," but used the name of a prominent writing system which emphasizes this type of movement. It is quite evident that this style of writing had not become so habitual with this pupil that it could be used without conscious effort.

THE PROBLEM.

This example may serve to illustrate the function which the scientific study of the handwriting movement has to perform. Its purpose may be thought of, first, as the examination of the current assumptions concerning methods of learning and of teaching. Its next task is to determine what the essentials of good writing are

apart from the presuppositions of current practice. The procedure which was followed in the experiment here described was to make an analysis of the behavior of good writers and poor writers, and to endeavor from this comparison to determine what characteristics are essential and what are negligible. The elements of the writing movement which were included in the study comprise position, the gross composition of the movement, and the detailed changes in the speed of the movement.

THE METHOD.

The method consisted of photographing the movement of the hand and arm by a kinesiographic camera. The camera was mounted directly above the hand of the writer and was speeded up to twenty-five exposures per second. This gave a separate photograph of the hand and of the preceding writing each twenty-fifth of a second. This record was sufficiently detailed for the purposes in hand.

The position of the hand at any point in the writing could be determined by putting the developed film in a projection machine, throwing the image on the screen with the machine still, and making a drawing from it. The nature of a gross movement could be determined by making a succession of these drawings to represent the positions at successive points in the writing. The most detailed examination of the movement was made by plotting the position of certain reference points on the hand at each successive exposure of the camera.

The speed changes in the movement were examined by recording the position of the pen point on the letter strokes at each successive exposure of the camera. From this record of the positions of the pen point at twenty-fifth of a second intervals, it was possible to determine the distance traveled by the pen point in successive units of time. These successive distances were plotted on a speed curve as is illustrated in Figures 3 and 4.

IMPORTANCE OF POSITION.

The methods of teaching handwriting in common use are similar to the methods of teaching other manual arts in that they assume that certain positions are most favorable to the successful prosecution

tion of the movement and attempt to teach the pupils to assume these positions habitually. These elements of position comprise a large part of the "good form" of the activity. The accepted features of position in writing have been very little changed for many years. The chief modifications within the last two or three decades have consisted in slight relaxation in the rigidity of the requirements.

Standard position illustrated in Figure 1 by the drawing from the hand of the late Mr. C. P. Zaner, one of the best known writing specialists and teachers. This position, however, deviates slightly from the rigid regulations already alluded to. The chief departure consists in turning the hand slightly over to the right. As a consequence of this the penholder points somewhat farther to the right than is prescribed. Furthermore, the penholder is allowed to rest in the depression between the thumb and forefinger, instead of crossing the knuckle joint at the base of the forefinger. The hand rests upon the nails of the third and fourth fingers, however, and the grasp of the pen corresponds quite accurately with the usual directions.

This fairly orthodox type of position may be compared with the hand position of another writer which departs much more widely from the position which is taught. See Figure 2. This is a good writer in the eighth grade. While the wrist is held fairly level, the pen is turned over to the right and the thumb and forefinger are drawn in more than would ordinarily be considered desirable. In spite of this, the writing of this girl is among the best to be found in the eighth grade. Other figures might be shown of individuals whose position is beyond criticism, but whose writing is very poor.

From a small number of such cases one might be tempted to conclude that position is of no importance whatever for good writing. A more comprehensive survey of a larger number of cases shows that position is of some value. The variations which appear among good writers, however, show that more deviation from the rigidity of the orthodox position may be permitted than is common. Our problem is to determine what the essential elements are and to distinguish them from the non-essential.

The more important characteristics of position may be gathered from an examination of Tables II and III. Table II is the key to the understanding of Table III. Hand position was first examined with reference to the extent to which the hand is turned down so that the wrist is level. This is called "degree of pronation." The method

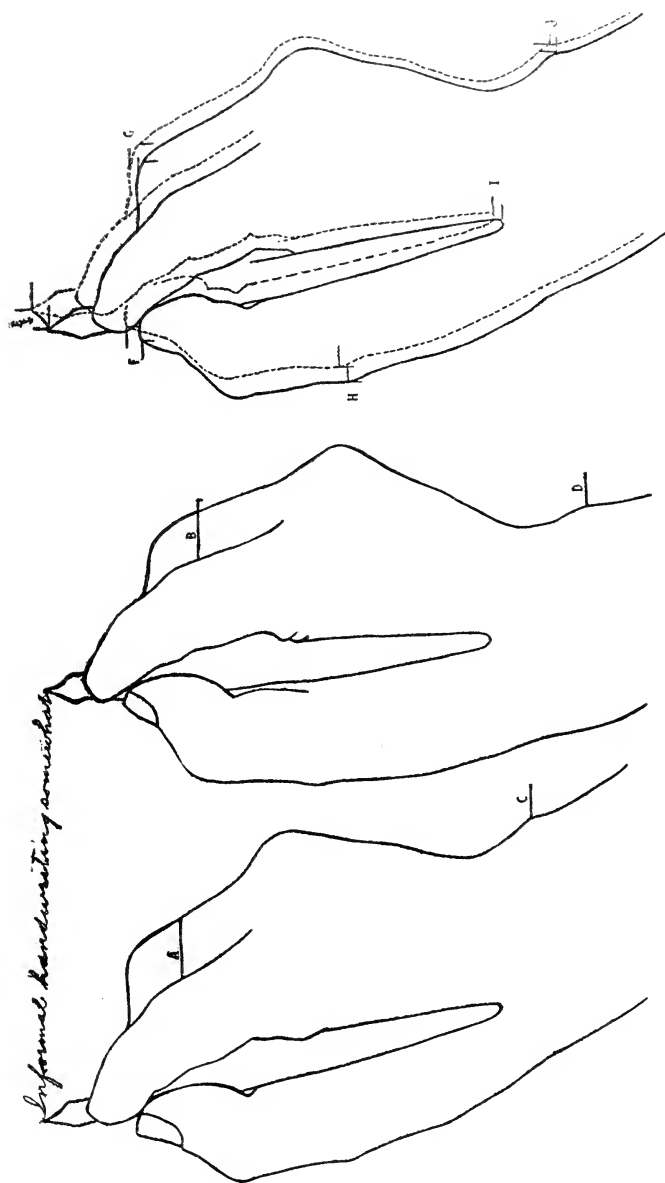


FIG. 1.—Illustration of the drawings showing hand position and gross movement.

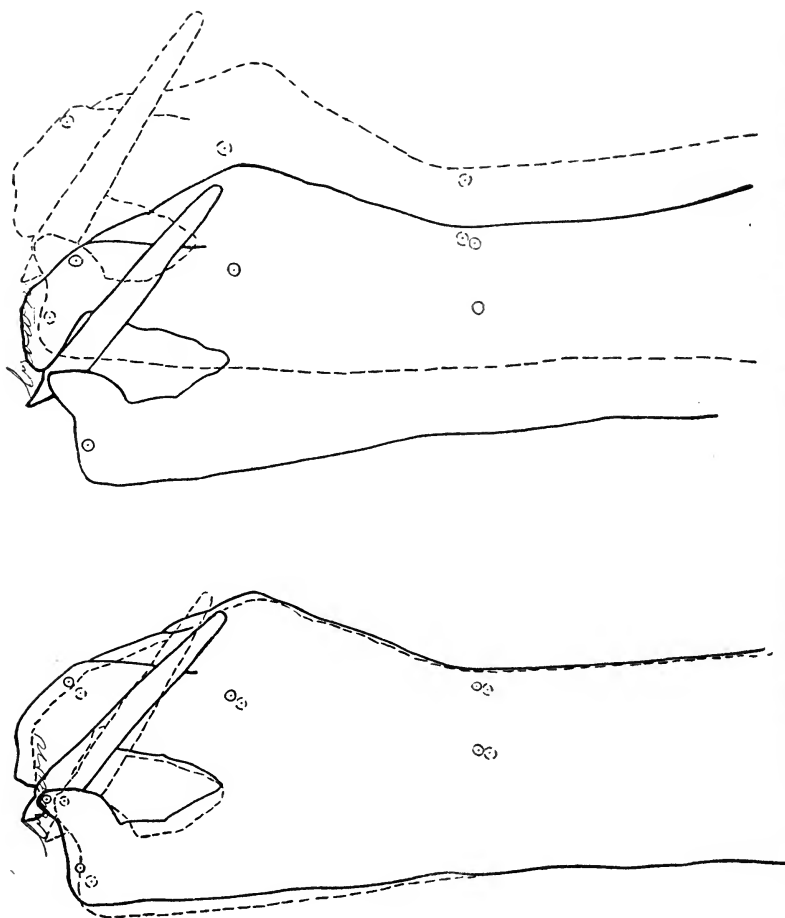


FIG. 2.—Hand position and gross movement of a good writer from Grade VIII

was to classify the individual writers into one of the groups defined in this table. The rest of the table is self-explanatory.

TABLE II.

Basis of Classification of the Individuals According to Their Hand Position.

A. Degree of pronation...	<ol style="list-style-type: none"> 1. Wrist level, or tilted less than 10°. 2. Wrist tilted 10° to 45°. 3. Wrist tilted 45° or more.
B. Support of the hand..	<ol style="list-style-type: none"> 1. On the third and fourth fingers. 2. On the side of the hand. 3. On the base of the hand.
C. Angle of the arm with the base line of the writing.	<ol style="list-style-type: none"> 1. Nearly perpendicular. 2. 10° to 30° from the perpendicular to the right. 3. 30° to 60° from the perpendicular to the right. 4. To the left of the perpendicular.
D. Relation of finger and thumb on penholder.	<ol style="list-style-type: none"> 1. Finger below thumb. 2. Finger opposite thumb. 3. Finger above thumb.
E. Looseness of grasp of penholder.	<ol style="list-style-type: none"> 1. Loose. 2. Medium. 3. Tight.

Table III indicates, for example, that two individuals of the poor writers in the University Elementary School held the hand so that the wrist was almost level. Ten of this group tilted the wrist from 10 degrees to 45 degrees and fourteen tilted it more than 45 degrees. The rest of the table is to be interpreted in a similar manner.

The degree of the pronation of the hand is influenced largely by training. This is shown by the marked change in the distribution of the poor writers in the University Elementary School after training. Whereas only one held the wrist level before training, twelve did so after training. A slightly higher percentage of the public school pupils than of the University children held the wrist level. This reflects the greater amount of formal drill which the public school children had received. Aside from the effect of training, there is little correspondence between the quality of writing and the degree of pronation of the hand. The chief correspondence with quality is found between the two groups of public school children.

Column B represents the manner in which the hand is supported. Here again the effect of training is manifest, both in the comparison of the University Elementary School group before and after training, and in the comparison of these children as a whole with the pub-

lic school children. There is also a slightly greater percentage of the good writers who support the hand upon the fingers than of the poor writers. The poor writers in greater number rest the hand upon the side or upon the base.

TABLE III.

Comparison of the Frequency of Various Hand Positions Among Several Groups of Children.

GROUP	Position	Pronation	Hand Support	Angle of Arm to Base Line	Relative Position of Thumb and Forefinger	Looseness of Grasp
		A No.	B No.	C No.	D No.	E No.
Univ. Elem. School: Poor writers (entire group).	1.....	2	8	5	10	4
	2.....	10	20	20	9	8
	3.....	14	..	2	3	10
	4.....	2
	5.....
Univ. Elem. School: Poor writers who took training.	1.....	1	5	1	8	2
	2.....	9	13	14	4	5
	3.....	8	..	2	2	7
	4.....
	5.....
Univ. Elem. School: Good writers.	1.....	2	7	6	9	7
	2.....	4	8	6	2	3
	3.....	6	..	2	..	1
	4.....	2
	5.....
Ray School: Poor writers.	1.....	1	5	2	7	4
	2.....	8	7	6	..	3
	3.....	3	..	2	1	1
	4.....	2
	5.....
Univ. Elem. School: Poor writers after training.	1.....	12	12	7	13	8
	2.....	6	5	13	1	6
	3.....	2	..	1
	4.....
	5.....
Ray School: Good writers.	1.....	4	8	5	8	4
	2.....	7	2	7	..	4
	3.....	1
	4.....
	5.....

The correspondence between the angle of the arm with the base line and the factors of training and of quality of writing is about the same as in the case of pronation. The effect of training is to reduce the numbers of cases in which the arm is held at an acute angle to the line of writing. Good writers, furthermore, hold the arm nearly perpendicular to the writing in somewhat larger numbers than do poor writers.

The relative position of the thumb and forefinger has a rather marked relation both to training and quality. The effect of training is seen to produce a natural position in which the forefinger rests on the penholder below the thumb. Good writers adopt this position more frequently than poor writers. Finally, a loose grasp is produced by training when it is directed to this aim, as in the case of the University Elementary School group; and in the case of the University children who had had relatively little formal drill the good writers exhibited much more looseness of grasp than the poor ones.

THE COMPOSITION OF THE WRITING MOVEMENT.

Composition refers to the make-up of the total writing coordination. One aspect of it deals with the relation between the arm and finger movement. The facts regarding the more significant phases of composition are presented in Tables IV and V. The interpretation of these tables is similar to that of Tables II and III.

TABLE IV.

Basis of Classification of Individuals According to Their Writing Movement.

A. Movement on letter strokes.	$\left\{ \begin{array}{l} 1. \text{ Movement of the whole arm.} \\ 2. \text{ Movement of the hand at wrist joint.} \\ 3. \text{ Movement of fingers mostly at third joints.} \\ 4. \text{ Movement of fingers mostly at second joints} \\ \quad \text{(ordinary finger movement).} \end{array} \right.$
B. Ease of movement along the line.	$\left\{ \begin{array}{l} 1. \text{ Easy.} \\ 2. \text{ Moderately easy.} \\ 3. \text{ Difficult.} \end{array} \right.$

TABLE V.

Comparison of the Frequency of Various Kinds of Movement Among Several Groups of Children and Adults.

Group.	Classes of Movement.	Movement on Letter Strokes.		Ease of Movement Along Line.	
		A. No.	B. No.		
Univ. Elem. School: Poor writers (entire group).	{ 1.....	6	1		
	{ 2.....	16	12		
	{ 3.....	4	9		
	{ 4.....	18	..		
	{ 5.....		
Univ. Elem. School: Poor writers who took training.	{ 1.....	3	..		
	{ 2.....	12	8		
	{ 3.....	2	6		
	{ 4.....	13	..		
	{ 5.....		
Univ. Elem. School: Good writers.	{ 1.....	3	5		
	{ 2.....	9	6		
	{ 3.....	3	1		
	{ 4.....	8	..		
	{ 5.....		
Univ. Elem. School: Poor writers after training.	{ 1.....	11	6		
	{ 2.....	6	7		
	{ 3.....	1	1		
	{ 4.....	2	..		
	{ 5.....		
Ray School: Poor writers.	{ 1.....	7	3		
	{ 2.....	1	6		
	{ 3.....	..	1		
	{ 4.....	7	..		
	{ 5.....		
Ray School: Good writers.	{ 1.....	5	8		
	{ 2.....	5	..		
	{ 3.....		
	{ 4.....	8	..		
	{ 5.....		
Adults: Poor writers.	{ 1.....	4	..		
	{ 2.....	1	..		
	{ 3.....	4	..		
	{ 4.....	8	..		
	{ 5.....		
Adults: Good writers.	{ 1.....	6	..		
	{ 2.....	2	..		
	{ 3.....	2	..		
	{ 4.....	6	..		
	{ 5.....		

Column A shows the proportion of children in the various groups who formed the letters by the movement of the whole arm, the movement of the hand, the movement of the fingers produced without bending them at the middle joint, and the ordinary finger movement, which consists chiefly in the hinge movement of the second joints.

We may examine chiefly the frequency of cases which fall under Heads 1 and 4 which comprise the arm movement and finger movement, respectively. There is no evidence in this table that good writers use the arm movement more commonly than poor writers. There is some evidence that training has an effect in producing arm movement, but the effect is not very pronounced. The most interesting fact is that the University group which took training developed a considerable amount of arm movement, in spite of that fact that this was not aimed at directly and was not mentioned in the instructions to the children. These instructions emphasized position, freedom of movement, an easy and fluent sideward movement of the hand, and a rhythmic swing in forming the letters. The training which followed from this type of emphasis apparently produced as an incidental result some arm movement. Undoubtedly the amount of arm movement which is thus directly produced, unlike that which is achieved by being consciously striven for, will remain a permanent element of his writing coordination.

Column B represents a characteristic which it was not easy to measure quantitatively. Pupils were grouped from an observation of the protographs of their writing movement, according as it appeared that they were able to carry the hand smoothly from left to right across the line while the letters were being formed. Ability to do this appears to be a characteristic of good writers as a class and also to be produced by training.

SPEED CHANGES OF THE WRITING MOVEMENT.

Probably more important than the position which is assumed in writing is its temporal character. The movement may proceed in a smooth, orderly, comparatively rhythmical fashion, or it may be jerky and irregular. The method of studying these changes has already been mentioned. Figure 3 shows the method in its successive stages. The top line represents the writing itself. The enlarged

reproduction of the word "brown" shows the record on which the position of the pen at the successive exposures of the camera is indicated. The numbers indicate that a succession of exposures took place while the pen remained at a given point. In other words, these represent pauses in the movement. The chart at the bottom shows the speed curve. The vertical units represent the distance which the pen traveled. The horizontal units represent successive twenty-fifths of a second. Each column shows the distance traveled by the pen in a given twenty-fifth of a second. The spaces on the base line at which there are no columns represent pauses. The strokes, which are written underneath the chart, show the parts of the letters which correspond to the part of the speed curve immediately above.

It is evident from an examination, either of the enlarged word or the speed curve, that the writing of this word consisted in a succession of fairly distinct strokes, every stroke separated by a slight pause from the preceding. It is also evident that the pauses in this case come at the natural turning points in the stroke, namely, those points in which the change in direction is marked. The speed on the long strokes is greater than that on the shorter strokes. The speed on a given stroke usually either increases gradually from the beginning to the end or is fairly even throughout the stroke. We have a general picture of regularity and of a correspondence between the length of the stroke and its speed.

This record may be contrasted with that of a poor writer shown in Figure 4. The movement is not divided into units corresponding to the natural units of the letters in the same fashion as in the preceding case. See, for example, the letter *b*. In the first place, there is a pause soon after the beginning of this stroke where we should not expect it. In the second place, the movement is not retarded at the top and the bottom of the long stroke as we should expect. Corresponding with this lack of retardation at the bottom, the stroke is very much rounded. Another illustration may be found in the letters *w* and *n*. There is usually a pause at the end of the last upward stroke of the letter *w*. Here the movement is simply retarded, but not sufficiently to make the sharp turning point, characteristic of this part of the letter. The connecting stroke and the first stroke of the *n* are made into a single stroke by obliterating the characteristic curves at this point. Corresponding to this modi-

A quick brown fox jumps over

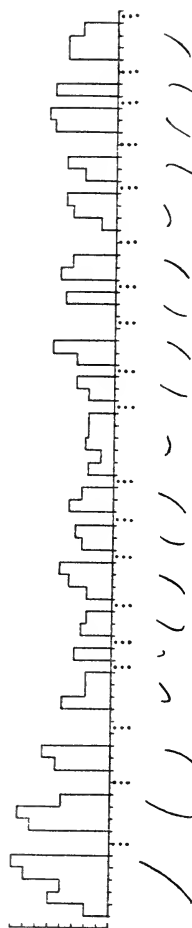
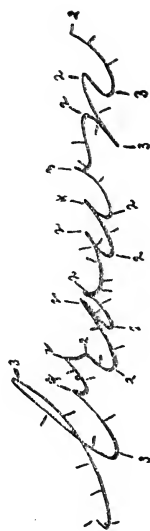


FIG. 3.—Record of a good writer from Grade VIII

a guide down for jumps over a logydog

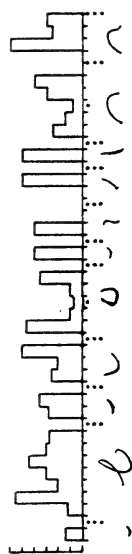


FIG. 4.—Record of a poor writer from Grade VII

fication is the continuity of speed with which this stroke is made. Another characteristic of this writing is that the greatest speed sometimes comes at the beginning or the end of a stroke rather than in the middle. Furthermore, a short stroke, such as the first or the second stroke of the *w*, may be made as rapidly as a much longer stroke. There is also to be found sudden alternation between pauses and rapid strokes. The movement, in other words, is irregular in its succession of strokes and of pauses, and it is not well adapted to the forms which are to be produced.

These illustrations may again be supplemented by the summary Table VI.

TABLE VI.

Percentage of Time Spent by Children in Pauses and at the End of Strokes.

	Grade.	Poorest.	Best.	Before Training.	After Training.
Univ. Elem. School....	VII-B.....	25	27
	VI-A.....	29	52
	VI-B.....	28	32
	V-A.....	16	30	16	34
	V-B.....	32	38	32	49
	IV-A.....	13	36	13	53
	IV-B.....	40	56	40	43
	III-A.....	37	39	37	53
	III-B.....	49	59	49	62
	II-A.....	53	58	53	60
	II-B.....	53	60	53	58
	Average.....	34	44	37	52
Ray School.	VIII-A.....	40	41		
	IV-A.....	32	42		
	Average.....	36	42		

The averages of this table show that the good writer divides the writing movement into a series of units, corresponding to the natural units of form of the letters, more radically than does the poor writer. The differences in the average of the percentages are marked and the difference is the same way in all of the groups. It therefore can hardly be accidental. The older writer and the poor writer both tend to make the successive strokes blend with one another. As practice makes the writer more skillful he is able to bring about a considerable increase in this continuity without the sacrifice of form. When the continuity is carried too far, however, it causes one form to run into another and therefore causes the writing to deteriorate.

SUMMARY OF CONCLUSIONS REGARDING POSITION AND MOVEMENT.

The evidence indicates that the following items of position have some relationship to excellence in writing:

The pronation of the hand to such a degree that the wrist is not tilted more than about 45 degrees from the horizontal;

The position of the forearm at an angle of about 90 degrees with the line of writing;

The support of the hand upon the third and fourth fingers rather than the side or the base of the hand;

A position of the thumb and forefinger on the penholder which is produced by slightly bending the joints. In this position the forefinger rests nearer the point than the thumb.

The conclusions regarding speed changes may be summarized as follows:

In good writing the continuity of the movement is broken up sufficiently to produce a retardation or pause at those points where there is a radical change in direction of movement.

The good writer adapts the speed to the strokes by making the longer strokes more rapidly than the shorter ones. The speed on similar strokes is similar.

The organization of the movement within strokes is ordinarily such that the speed is greater in the middle of a stroke and less at the beginning and end.

There is ordinarily not a sharp contrast between the speed of successive parts of the writing, such as *a* would be represented by a sudden, jerky movement.

EXPERIMENT WITH THE TRAINING CLASS.

The susceptibility of some of these characteristics of good writing to training has already been discussed in connection with a group of the University Elementary School pupils. A more extensive experiment of the applicability of such principles, taken as a whole, was made with a school in Kansas City, Kansas, for a period of a year. A system of exercises was worked out in connection with the training class in the University Elementary School. These exercises were mimeographed and taught to the pupils in Kansas City. The progress made by the pupils of this school was compared with the progress of the pupils in two other schools in the same city. The results are summarized in Table VII.

TABLE VII.

Comparison of Average Progress of the Training School (Longfellow) and the Two Check Schools.

Grade and School.	Speed ²	Form ³
Grade II:		
Standard.....
Longfellow average.....	29.3	4.0
Bryant average.....	30.9	2.0
Quindaro average.....	19.9	2.2
Grade III:		
Standard.....	13.2	1.4
Longfellow average.....	25.8	3.4
Bryant average.....	29.7	-1.2
Quindaro average.....	18.5	-1.0
Grade IV.		
Standard.....	7.4	2.0
Longfellow average.....	28.1	3.2
Bryant average.....	30.4	-0.7
Quindaro average.....	38.4	-0.7
Grade V:		
Standard.....	7.9	1.4
Longfellow average.....	28.4	3.5
Bryant average.....	26.7	-0.2
Quindaro average.....	17.4	0.9
Grade VI:		
Standard.....	3.7	2.1
Longfellow average.....	48.6	2.6
Bryant average.....	11.1	1.5
Quindaro average.....	-12.1	2.2

²In letters per minute.

³In units on the Freeman scale.

Grade VII:

Standard.....	5.1	1.8
Longfellow average.....	15.3	7.7
Bryant average.....	12.3	2.3
Quindaro average.....	6.4	-1.0

Grade VIII:

Standard.....	5.1	1.1
Longfellow average.....	20.6	6.0
Bryant average.....	43.8	-0.5
Quindaro average.....	13.1	0.2

Average Gain:

Standard.....	7.1	1.6
Training school.....	28.0	4.34
Two check schools.....	20.5	0.43

It is of course impossible to say what features of the method were responsible for the gain which is shown here. The training school gained twenty-eight letters per minute in speed and 4.34 points in form, while the two comparison schools gained 20.5 letters per minute in speed and .43 points in form. The method of producing an improvement in position need not be described. The organization in speed of the writing movement was effected by the application of counting to the writing of the individual letters and of combinations of letters and words. This counting was applied almost solely to the writing of actual letters themselves and not, as is customary, to the writing of exercises. Experience seems to indicate that counting on exercises has a rather remote effect upon the organization of the speed in writing the letters themselves. This investigation, then, confirms, to some degree at least, the conclusions that were drawn from the earlier experimental work regarding the features of position and movement which are particularly important for good writing.

INTELLIGENCE AND ITS MEASUREMENT: A SYMPOSIUM*

XIV. By B. R. BUCKINGHAM.

The nature of intelligence and its measurement. Doctors differ in defining intelligence, yet even the man in the street will pass judgment upon the actions of his fellows as either intelligent or stupid. It seems to me that for practical purposes we may get along very well without a definition of the central quality in virtue of which behavior is effective or lacking in effectiveness. This is fortunate because it seems unlikely that any formulation of the nature of this quality will ever be so commanding as to win the assent of a majority even of those who are qualified to pass judgment. We may, it is true, be making virtue of necessity in thus deprecating a description of mentality in the abstract. But even if general intelligence as a personal possession could be defined without reference to its manifestations, the most useful definition would still be in terms of behavior. One who subscribes to this doctrine may occasionally experience a qualm of ingrowing consciousness as he realizes the possibility of thus being dubbed a behaviorist. But perhaps some of us have been called worse names; and aside from the instinctive aversion to being tagged, we shall make shift to endure the epithet with equanimity.

Let us seek, therefore, the nature of intelligence not in some central capacity such as association (Ebbinghaus) or attention (Wundt), but in the nature of acts which the central capacity, whatever it may be, makes possible. The question then becomes one which we can talk about with some chance of being understood. We shall not often nor widely disagree as to whether or not an act in a given situation is effective. Having observed the behavior of a person in a number of similar situations, we rather easily—too easily sometimes—conclude that he has a certain amount of the special ability or aptitude which belongs to the particular type of situation. Again, when we observe a person's behavior, not only with reference to similar situations, but also with reference to a variety of different

*Editorial note: We are glad to print this month Dr. Buckingham's statement, which came to hand too late for the April issue. Further interpretation of the material in this symposium will appear in the September issue.

situations, we are justified in concluding that he possesses, to a certain degree in each case, a number of special abilities or aptitudes appropriate to these situations. Finally, with that insistent habit of generalizing which we undoubtedly possess—a habit in virtue of which we seek to reduce diversity to a type, and to express variety in terms of an average—we deduce from the special abilities a general ability. If a person is effective in his behavior in a variety of situations, we assert his general superiority and we ascribe this general superiority to that central directive power which we term intelligence.

It appears that the two essentials in the development of intellectual ability (or the ability to act effectively under given conditions) are, first, a recording mechanism, and, second, something to record. When the late Walter Bagehot characterized the ability of Shakespeare as the product of a first-rate imagination working upon a first-rate experience, he gave, in my judgment, all that was essential.

These, then, are the two variables which enter into both the nature of intelligence and its measurement—the experience and the experiencing nature. If we can control either of these variables, we can measure the other. The experience of school children is reasonably under control. The measurement of the intelligence of school children becomes, therefore, a problem whose solution falls within the bounds of possibility. To my mind, the measurement of the intelligence of persons about whose experience we know little or nothing is highly unreliable. Fortunately, it is with school children that our chief interest lies. The large question which we wish to determine with respect to these children is their educability, or their ability to learn. Whatever may be the theoretical conception of intelligence, it is submitted that it exists for a purpose and must eventuate in something. Pupils who have been brought under the influence of a school regimen will learn somewhat in proportion to their intelligence. A measure, therefore, either of the rate at which learning takes place or of typical products of learning will constitute a measure of intelligence.

Not inappropriately, therefore, do our intelligence tests include tests of learning, such as the digit-symbol test. But by far the greater part of the intelligence tests measure in some way the product of previous learning—i e., they measure knowledge or informa-

tion in a broad sense. This is obviously true, for example, among the verbal tests. It is certain that success in solving analogies depends quite as much upon knowledge of the terms used as upon any innate capacity to perceive relationships. The same may be said of success in picking out synonyms or antonyms, in completing sentences, or in matching proverbs. Success in non-verbal tests likewise depends upon knowledge; or, in other words, upon the product of experience in an experiencing nature. The completion of a picture is impossible unless one knows what the picture ought to be. Similarly, the crossing of a diagram which does not belong with others involves knowledge about the diagrams in question. Even success in following directions requires acquaintance with the terms used and with certain instruments and visual symbols. The non-verbal test, therefore, does not differ essentially from the verbal test except in the kind of the knowledge or experience required to negotiate it.

Accordingly, it seems to me that whatever definition we may give to intelligence in the abstract, we are justified from an educational point of view in regarding it as ability to learn, and as measured by the extent to which learning has taken place or may take place. Of course, in seeking to measure the extent to which learning has already taken place, we do well to avoid specific sorts of learning. Thus, a test of the ability to answer questions in geography would be an inferior measure of intelligence, because it would depend to a large degree upon the amount and recency of a particular type of instruction. There are, however, certain products of intelligence which are of so general a nature that we may appeal to them with some confidence. These we attempt to include in our mental tests in the form of tasks to be performed. The quest for the right kind of tasks is still going on. It is like searching for a few master keys that will unlock many doors.

Next steps in research. I should distinguish at least eight important lines along which investigations should proceed in the measurement of intelligence.

1. The reliability of our present instruments of measurement should be more satisfactorily determined. By reliability I mean the extent to which the scores obtained in the tests are consistent with those obtained by repeating the test with the same individual and under the same conditions (allowing for the effect of practice). When, for example, we infer that a pupil's mental age is ten years,

we wish to know the Probable Error of the result. The public has a right to know the extent to which it may rely upon our determinations.

2. There is at least as great a need for determining the validity of intelligence tests. By validity I mean the extent to which they measure what they purport to measure. If for educational purposes we define intelligence as the ability to learn, the validity of an intelligence test is the extent to which it measures ability to learn. In a very real sense, validity is more important than reliability. No one, for instance, is interested in the consistency of the results of a test which fails to measure the thing it is designed to measure. Such a test would merely be consistently valueless.

3. Most tests consist of a number of sub-tests or tasks each involving a certain type of response. The analysis of the tasks which are now in use and the search for new ones with a view to utilizing a few which are highly symptomatic of intelligence is one of the next steps in research. Here the investigation may proceed along two lines: first, the analysis of intelligence, or, more concretely, of intelligent behavior; and second, the statistical treatment of the results of this analysis through correlation and partial correlation. By the second process the most highly significant of the tasks previously suggested by the analysis of intelligence will be preserved, while at the same time in formulating a battery of tasks duplication will be avoided. The task which correlates most highly with a criterion of general intelligence is one which we wish to use. But if it correlates highly with another task equally related to general intelligence, we shall not use both these tasks in the same battery because each would tell the same story.

4. Just now we are greatly interested in the constancy of the Intelligence Quotient. The bearing of this on prophecy and on long-term planning in school administration is obvious.

5. In addition to instruments measuring general ability, a large number of instruments for measuring specific abilities are desirable. These will have particular value in educational and vocational guidance. A creditable beginning has been made in devising such instruments. "Next steps" should contemplate moving further in this direction.

6. It is generally agreed that intelligence among children and young people differs at different levels of development, not only in

degree, but also in kind. Accordingly, a single mental test cannot be successfully used over a wide range of intellectual levels. As a product of future investigation we hope for a hierarchy of tests, each appropriate to a given level, and each linked with the others by ascertained relationships. Thus for certain purposes we may be able to transmute scores obtained on one test into the scores of another test.

7. It is perhaps a question whether moral qualities and qualities of the "will" should be considered in this connection. It is, however, so clear that success both in school and in life depends upon these qualities, that we cannot avoid pointing out the need which we are under of supplementing our measures of intelligence with measures of the non-intellectual traits which condition human success.

8. Investigators ought seriously to consider the questions of standard procedure and standard form in regard to tests. The simplicity and fullness of directions to examinees, the provision of fore-exercises, the arrangement of material on the printed page, the provision of facilities for checking papers, the directions for handling the scores, and the amount of statistical labor to be required—these are some of the kinds of standardization which we shall doubtless obtain as new tests are devised and the law of the survival of the fittest comes into play.

THE TRUE-FALSE TEST AS A MEASURE OF ACHIEVEMENT IN COLLEGE COURSES

ARTHUR I. GATES,

Teachers College, Columbia University.

Every teacher is painfully aware of the enormous demands upon time and energy required to read and evaluate sufficient examination papers to secure a representative measure of achievement. When classes are many and large, the task becomes a major feature of the "teaching burden" and relief is frequently sought by turning over part of the work to "readers" or assistants. That the grading of examination even by teachers of extensive experience is grossly unreliable is indicated by numerous studies carried out within the last decade. The investigations of Cattell, M. F. Meyer, Foster, Finkelstein, Kelley, Starch, Starch and Elliott and others are too well known, however, to the readers of this Journal to necessitate summary here.

In a limited number of fields, reading, writing, spelling, arithmetic, and the fundamental subjects, generally, standardized tests, developed by means of the technique of mental measurements, are available, but in the content subjects of secondary schools and colleges, such tests are few. The unreliability of the conventional essay examination is suspected, but adequate substitutes are wanting.

During the past three years, different forms of examinations patterned after the forms of intelligence and educational tests have been used, together with the conventional essays, and in some cases with tests of intelligence in graduate and undergraduate courses in educational psychology at Teachers College. Sufficient data are now available to make it possible to present a reasonably reliable report of the usefulness of some of these tests.

The Subjects.

Four classes (20, 32, 72, and 74 students, respectively) of an undergraduate course in educational psychology.

Six classes (57, 66, 68, 74, 80, and 82 students, respectively) of graduate courses in intermediate educational psychology. The numbers indicate the total population completing all of the tests upon which the computations are based.

The Tests.

1) Mid-term, or final 1 and 2 hour examinations of the conventional essay type, graded carefully by an instructor, or an assistant, or independently by an assistant and an instructor, or by two instructors.

2) Home-written work; abstracts of books or articles, reports of experiments, essays or special topics, or problems, etc., graded by the instructor, an assistant, or by each, independently. The scores used in correlation are the scores on all such work for a half-semester, representing an average of approximately 8 hours actual writing.

3) The Army Alpha or Thorndike Intelligence Test (Part I or Parts I and II).

4) The new examinations, mostly of the true-false type. The test consists of a series of 30 or more statements which the student checks as true or false, or in some way, such as by underlining, indicates the correct one of two possible answers. The following are sample questions:

A. Questions concerning information given in text or lectures. (Answer "True" or "False.")

1. The Auditory area is in the Temporal Lobe of the cortex.
2. According to Terman's findings, the average intelligence of boys is markedly superior to the intelligence of girls.
3. In learning poetry, the "whole" method is usually found to be superior to the "part" method.

B. Questions concerning information less explicitly given.

1. The curve of forgetting has been more thoroughly investigated than has the curve of learning.
2. If the cerebral hemispheres of a frog were removed, it would remain completely paralysed (assuming it survived the operation).
3. The general laws of learning indicate the wisdom of using "water-wings" in learning to swim.
4. If we find a large percentage of mentally inferior children in the slums, it proves that slum environment is one cause of mental deficiency.

The statements or questions are presented in the following ways: 1) *Oral Presentation*. The instructor reads the question one, two or more times. He may change the wording to make the meaning clear. A little less than one minute per question is sufficient. 2) *Oral and Visual Method*. The instructor reads the question, writes it on the board and reads it again. A trifle more than a minute per question is required. 3) *Visual Method*. Each student is given a mimeographed copy of the questions. A time allowance of one minute per question is sufficient. The time allowances are in all cases generous; in fact, the stress of hurrying is carefully avoided.

Usually the students correct their own results, as the instructor rereads the question and gives the proper answer. The score equals total number of questions minus twice the number wrong. Since this method of scoring is well known and has been defended elsewhere,* it will not be discussed here. Finally the scores are displayed on the board, showing each student his position in the group. A period of 50 minutes is sufficient to give, score, tabulate and discuss the results of a test of 40 questions. Other considerations of the true-false type of examination will be presented in the last section of this paper.

5) *The Criterion*. In constructing the criterion, the sum of all essay examinations was given the weight of 1.0, the sum of the True-False Tests 1.0, the sum of the written work 0.5; the sum of class recitations, oral quizzes, special conferences, etc., 0.5.

Methods of Computing Results.

All coefficients of correlation presented in the following tables were computed by the Pearson Product-Moment formula.

The several coefficients falling within a decade are presented in a line, thus giving a rough picture of the distributions. The number of cases, the average, the S. D. (σ) of the dist., and the σ (true-obt. av.) are given for each array. In no case has a coefficient been corrected for attenuation for the reason that we are not interested in ideal relationships, but in those which obtain in the existing data. All coefficients listed are positive unless otherwise indicated.

*McCall, W. A. *Journal of Educational Research*, 1920, 1, 33-46.

TABLE I.

Correlations of one essay examination with another.

13, 15,			
21, 21, 23,	25, 25, 26, 29		
33, 38, 39			
42, 47, 49			
50, 52, 58			
65			
N = 19	Average = .35	S. D. = .146	S. D. (tr-obt. av.) = .033

TABLE II.

Correlations of one measure of written work with another.

21, 26			
32, 33, 35, 37			
41, 41, 42, 45, 46			
51,			
60, 62, 66			
N = 15	Average = .43	S. D. = .125	S. D. (tr-obt. av.) = .034

TABLE III.

Correlations of one True-False test with another.

13, 19			
21, 24			
31, 33, 35, 37, 37, 39			
41, 41, 43, 45, 45, 45, 46, 46, 47, 48, 48, 48, 48			
50, 50, 51, 53, 53, 53, 55, 55, 55, 55, 56, 57, 58, 58, 58, 59			
60, 60, 61, 62, 64, 64, 65, 65, 67, 69			
71, 73, 73, 75, 78, 78, 78			
80, 81, 81			
N = 59	Average = .54	S. D. = .152	S. D. (tr-obt. av.) = .02

TABLE IV.

Correlations of one True-False with one essay examination.

-11			
-02			
03, 08			
11, 11, 13, 15, 18, 19			
20, 20, 21, 21, 23, 25, 28			
31, 34, 37, 38,			
42, 44, 45			
51, 57, 59			
61, 64, 68			
71, 74			
84			
N = 33	Average = .33	S. D. = .213	S. D. (tr-obt. av.) = .04

TABLE V.

Correlation of one True-False with written work.

00, 04			
11, 14, 17			
21, 22, 25			
30, 33, 37, 38			
41			
64			
N = 14	Average = .25	S. D. = .144	S. D. (tr-obt. av.) = .04

TABLE VI.

Correlations of an essay examination with written work.

15, 23, 26, 31, 53 N = 5 Aver. = .30 S. D. = .12 S. D. (tr-obt. av.) = .05

Inter-correlations of essays, written work and True-False Tests.

The inter-correlations of the essay examinations are decidedly low, for reasons that cannot be discovered from the available data. From other investigations it appears that personal factors involved in the gradings are, in part, responsible. That such may be the case is indicated by the following correlations obtained from the estimates of various sets of essays, each set graded by two judges independently: 27, 37, 41, 46, 50, 52, 66, 74, 90. Average = .54, S. D. = .183. The correlations may be low when the judges accurately estimate different features of the essays which are not positively correlated or they may be low because the judges inaccurately estimate the same features. Many other possibilities make speculations futile. The correlations of estimates of sets of written work, each by two judges, are similarly scattered: 17, 25, 29, 33, 41, 49, 63. Average = .36, S. D. = .144.

The average of the inter-correlations of the True-False tests are higher than those obtained with the essay examinations or the written work; 90 per cent of the coefficients reaching or exceeding the median of the former and 80 per cent reaching or exceeding the median of the latter. It should be recalled that the True-False test represents about 30 minutes' work, whereas the essay requires an hour and the written work about 8 hours. High inter-correlations would not, however, necessarily indicate that the tests are reliable measures of achievement, since they may be produced by many other factors, e. g., capacity to adjust to this particular sort of work. The dispersion of True-False inter-correlations is great. One might speculate about this, or study the conditions under which high and low coefficients were produced (as has been done to some extent), but without illuminating results. Whether the low correlations here found are due to badly selected series of questions, or to low correlations between the different contents tested, each content being accurately tested, to inequalities of previous equipment, to individual variability of performance, or to other causes, is not yet determined. The significance of the inter-correlations will depend upon the correlations of the tests with other tests and the criterion.

A single True-False test gives a correlation with an essay examination which is low, but as high as the average inter-correlation of essay examinations. The correlation of written work with the essay is almost as high, while the correlation of true-false with written work is slightly less. The σ (tr-obt. av.) indicates that real differences between these coefficients are improbable. The True-False test, then, yields as high a correlation with an essay examination or a half year of written work, as would another essay or half year of written work. If it should develop that the True-False tests yield as high a correlation with the criterion as either other test, it would be sensible to use them as *measures of achievement* because of the great saving of time.

TABLE VII.

Correlations of essay examination with criterion of achievement.

27, 28			
32, 38			
42, 48			
51, 51, 53			
63, 68			
70, 71, 73, 76			
82			
90			
N = 17	Average = .56	S. D. = .187	S. D. (tr-obt. av.) = .045

TABLE VIII.

Correlations of written work with criterion.

13,			
25, 27			
32, 38, 38			
40, 41, 42, 45, 46			
53,			
61			
N = 13	Average = .39	S. D. = .119	S. D. (tr-obt. av.) = .03

TABLE IX.

Correlations of True-False with Criterion.

32, 39			
45, 45, 46, 46			
50, 51, 51, 52, 52, 52, 53, 56.			
61, 63, 64, 64, 65, 67, 67, 68, 68			
71, 73, 73, 74, 75, 75, 77, 77, 78			
83, 84, 85, 86, 88			
90, 91			
N = 39.	Average = .65	S. D. = .151	S. D. (tr-obt. av.) = .026

TABLE X.

Correlations of 2 and more True-False Tests with the criterion.
The first row of horizontal figures indicates a class.

	1	2	3	4	5	6	7	8	9	10	Aver.	S. D.
2 T. F. Tests	.46	.52	.56	.59	.65	.69	.72	.82	.84	.86	.672	.127
3	52	59	58	62	69	74	73	84	84	88	.701	.118
4	58	60	63	66	75	79	79	86	88	90	.744	.113
5	68	68	68		83	83	85	86	90	91	.803	.090

TABLE XI.

Correlations of 2 and more essay examinations with the criterion.

	1	2	3	4	5	6	7	8	9	10	Aver.	S. D.
2 essay examinations	40	43	47	49	54	58	60	62	62	66	.541	.085
3 essay examinations	38	52	49	53	56	57	58	65	60	68	.556	.081
4 essay examinations	49		47	49	58	61	58	65	63	66	.573	.063
5 essay examinations			53	60	60	60			64	68	.610	.046

Correlations with the Criterion of Achievement and with Intelligence Ratings.

In constructing the criterion, it will be recalled that the sum of the essays were weighted 1; the sum of the True-False 1, the sum of written exercises, etc., 0.5; class recitations and all other information 0.5.

The correlation of the written work with the criterion is lowest. This may be due in part to the smaller weight given it in the criterion. We are mainly interested in True-False and essay coefficients, and it would appear from Tables V and VI that any weight given the written work would influence these about equally. That the average correlation + 0.65 of True-False with criterion is greater than + 0.56 of essay with criterion is evidenced by the σ (tr-obt. avs.). The difference is $.09 \pm \sigma$ (tr-obt. Diff.) .052.

The True-False test thus appears, all things considered, to be the most reliable single measure of achievement. What will be the results when several tests of the same type are combined? How many tests must be given to yield a satisfactory correlation with actual achievement? Tables X and XI give the correlations of 2 to 5 tests with the criterion. The fact that increasing the number of

True-False tests up to 5 brings rather large and uniform increases in the correlations, which is less true of the essay examination, is of marked importance. A group of 3 or 4 True-False tests clearly gives a higher correlation with the criterion than does an equal number of essay examinations.

Just what number of tests is considered "satisfactory" will depend upon our notion of a "satisfactory" correlation and our confidence in the criterion here used. In the course of time a number of cases of sharp disagreement between the scores secured by individuals in the various types of tests have occurred, and of these a number have been further tested as completely as circumstances would permit. The following case may be cited as typical. A student, ranking in the lowest tenth in four True-False examinations, was graded in the upper tenth in an essay examination by an assistant and in the second tenth in another such examination by the instructor. As a result of class-room questioning, two special examinations and a private oral quiz, we were convinced that the True-False results gave the proper measure of achievement. The student was dull but diligent; much material virtually had been memorized by rote. This material, rewritten with excellent penmanship and correct grammar, gave the impression of general mastery of the subject. Ability to apply the facts and principles was unusually deficient. Since nearly all of the special cases turned out similarly, the implication is that the better the criterion of achievement, the greater the predictive value of True-False as compared to the essay. Experience with the True-False examination has convinced us that it affords an opportunity to test more adequately the power to apply and use information in solving a concrete problem or meeting a novel situation.

The correlations with intelligence ratings will be of interest in this connection. Tables XII, XIII and XIV show average correlations with intelligence as follows: A single True-False + .406, an essay + .344 and written work + .255. The difference between True-False and essay is + .062 which is about the same as the σ (tr.-obt. diff.) = $\pm .067$. From tables XV, XVI and XVII it appears that combining several essay examinations makes no noticeable increase in the correlation, whereas the addition of True-False tests yields a steadily increasing correlation. A team of 3 or 4 True-False tests gives a clearly higher coefficient than a similar team of essays.

TABLE XII.

Correlations of Written Work, with Intelligence Ratings.

08				
17, 18				
22, 25, 26				
34, 39				
41				
N = 9	Average = .255	S. D. = .102	S. D. (tr.-obt. av.) = .034	

TABLE XIII.

Correlations of essay exam. with Intelligence.

06				
13, 15, 19, 19				
25, 29, 29				
32, 35, 36, 37				
41, 47				
52, 56				
61				
72				
N = 18	Average = .344	S. D. = .170	S. D. (tr.-obt. av.) = .04	

TABLE XIV.

Correlations of True-False with Intelligence.

13, 15				
21, 24, 29, 29				
32, 36, 39, 39, 39				
41, 45, 45, 46, 47, 49				
50, 56				
60, 66,				
71				
N = 22	Average = .406	S. D. = .15	S. D. (tr.-obt. av.) = .032	

TABLE XV.

Correlation of the two Half-Term measures of written work with Intelligence, for four classes.

17, 24, 30, 38,	Average = .272
-----------------	----------------

TABLE XVI.

Correlation of 2 or more essay examinations with Intelligence, arranged by classes.

	1	2	3	4	5	Aver.	S. D.
2 essays	24	29	35	39	43	.34	.052
3 essays	26	28	39	39	40	.344	.057
4 essays	30	29		40	41	.350	.054

TABLE XVII.

Correlations of 2 or more True-False with Intelligence arranged by classes.

						Aver.	S. D.
2 True-False	33	38	45	49	51	.432	.067
3 True-False	38	44	43	54	57	.472	.071
4 True-False	41	47	52	59	55	.508	.063
5 True-False	45	52		63	58	.545	.063

That the True-False test is not merely a new test of general intelligence is indicated by the fact that it yields higher correlations with the criterion of achievement; correlations which are higher than those obtained by the conventional examination. It would be useful to compare the correlations yielded with intelligence and with achievement by True-False tests composed of questions of information with questions requiring applications or problem solution, but it cannot be readily accomplished with the available data.

Correlations Obtained by Different Methods of Presentation.

The correlations of single tests with the criterion given in Table IX have been classified according to the method of presentation with the following results:

- 1) Oral presentation, $n = 18$, average = $.65 \sigma$ (tr.-obt. av.) $\pm .043$.
- 2) Oral and visual (written on board), $n = 8$, average = $.62$, σ (tr.-obt. av.) = $.048$.
- 3) Visual (mimeographed), $n = 13$, average = $.66$, σ (tr.-obt. av.) = $\pm .030$.

While the data are too few to make generalization secure, no method of presentation shows superiority.

Summary.

The value of a True-False examination as a measure of achievement in college courses is demonstrated by the following comparisons:

1. It yields correlations with other tests of the same type, averaging $+.54$, σ (tr.-obt. av.) $\pm .02$, as compared to inter-correlations of essay examinations of $+.35$, σ (tr.-obt. av.) $\pm .033$, and inter-correlations of written work of $+.43$, σ (tr.-obt. av.) $\pm .034$.

2. The True-False tests yield a correlation of $+.33$; σ (tr.-obt. av.) $\pm .04$ with an essay examination, which is as high as inter-correlations of essay examination.

3. The True-False test yields a correlation with written work of $+.25$, σ (tr.-obt. av.) $\pm .04$, which is probably as high as the correlation of essay examination with written work, $+.30$, σ (tr.-obt. av.) $\pm .05$.

4. The correlation of a True-False test with the criterion of achievement is $+ .65$, σ (tr.-obt. av.) $\pm .026$, as compared to $+ .56$, σ (tr.-obt. av.) $\pm .045$, for the essay examination and $+ .39$, σ (tr.-obt. av.) $\pm .03$ for the written work.

5. Combining several True-False tests produces markedly higher correlations with the criterion which is not true of the essay examinations.

6. Correlations with intelligence tests are: for the True-False $+ .406$, σ (tr.-obt. av.) $\pm .032$, for the essay examinations, $+ .344$, σ (tr.-obt. av.) $\pm .04$ for the written work, $+ .255$, σ (tr.-obt. av.) $\pm .034$.

7. Combining several True-False tests produces a marked increase in the correlations with intelligence, which is not true of the essay examinations.

8. Intensive study of certain cases indicates that the better the criterion of achievement, the greater the predictive value of True-False as compared to the essay examination.

Other Advantages of the True-False Examinations.

The following advantages of the True-False examination are not derived from objective data, but are accumulations of opinions generally voiced by those who have given or those who have taken such tests.

1. It saves an enormous amount of time. All the work of scoring is saved, since it is instructive to have the students correct the papers.

2. It is possible to develop standards of achievement by which one class may be compared with others.

3. Since each test includes 30 or more questions, it is possible to examine the field much more thoroughly than with the conventional examinations.

4. More refined units of measurement are secured; the distribution is wider than can be reliably obtained by the conventional examination.

5. Students report that the True-False examination is conducive to more effective methods of study. They study to understand and apply, rather than to commit to memory for purposes of sheer reproduction.

6. The True-False examination is really an excellent teaching instrument. By correcting the answers at once, haziness, misunderstandings and ignorance of facts and principles are clearly shown.

7. Showing the student in position in the group by distributing the scores on the blackboard is of recognized pedagogical value.

8. According to votes, at least 90 per cent of the students express a preference for this type of test. Some of the reasons frequently given are:

a) When the hour is over, the results are known; there are no days of uncertainty to be suffered.

b) After the initial adjustment, less nervousness before and during the examination is felt, usually.

c) Anticipation of being misunderstood, of favoritism or the opposite, is unnecessary.

d) The work during the examination is less exhausting. There is little writing or eye strain; no need of fear of being unable to finish within the time limit.

TRANSMUTATION OF VALUES ON THE THORNDIKE AND AYRES HANDWRITING SCALES: A CORRECTION

T. L. KELLEY,
Leland Stanford Junior University.

My attention has recently been called to a numerical error which occurred in a study of mine appearing in *THE JOURNAL OF EDUCATIONAL PSYCHOLOGY* of December, 1914, entitled "Comparable Measures." I take this opportunity to resubmit conclusions based upon correct calculations. This article dealt with the comparability of scores upon the Ayres and Thorndike Handwriting Scales, and gave an equating of the two scores, which is in error. The constant 2.917 reported on page 593 should be 2.223. The constant 7.90 derived from the inaccurate value 2.917 should be 7.147. Other values derived from these are in error. I list below the materially inaccurate statements and following them the revised correct statements.

Inaccurate statement: "The average variation of the estimate (of handwriting) upon the Ayres scale is 7.40 and upon the Thorndike scale when reduced to comparable units 8.68, which gives a difference of 1.28 in favor of the Ayres scale. The probable error of this difference is unknown, but is greater than .372."

Correct statement: "The average variation of the estimate (of handwriting) upon the Ayres scale is 7.40 and upon the Thorndike scale when reduced to comparable units 7.85, which gives a difference of .44 in favor of the Ayres scale. The probable error of this difference is unknown, but is greater than .372."

Incorrect Statement. Equal Measures.

Ayres Scale.	Thorndike Scale.
9.5	5
17.4	6
20	6.33
25.3	7
30	7.60
33.2	8
40	8.86
41.1	9
49.0	10
50	10.13
56.9	11
60	11.39
64.8	12
70	12.66
72.7	13
80	13.93
80.6	14
88.5	15
90	15.19
96.4	16

Correct Statement. Equal Measures.

Ayres Scale.	Thorndike Scale.
13.3	5
20	5.96
20.5	6
27.6	7
30	7.34
34.7	8
40	8.74
41.9	9
49.0	10
50	10.13
56.2	11
60	11.53
63.3	12
70	12.93
70.5	13
77.6	14
80	14.33
84.8	15
90	15.73
91.9	16

DEPARTMENT FOR DISCUSSION OF RESEARCH PROBLEMS



Conducted by LAURA ZIRBES



This department has a two-fold function. It aims to serve research workers as well as educators, whose work brings them in close contact with children in the schools. It hopes to accomplish this service by suggesting research studies, which will meet well-defined school needs.

In order that this service may be real and effective, the co-operation of research workers and school people is desired. Correspondence with reference to the following questions will be considered in selecting topics for future discussions.

- a. Which of the studies proposed would help you to solve a practical problem?
- b. What topics might well be added to this list? Replies may be addressed to: Miss Laura Zirbes, 646 Park Ave., New York City.

SOME THINGS I WANT TO DO OR SEE OTHERS DO

WM. A. McCALL,
Teachers College, Columbia University.

1. *Technique for Constructing Validation Criteria.* There are many steps in the process of scale construction. Among these the following may be mentioned: (a) Preparation and organization of test material, (b) Preparation of instructions for applying and scoring the test, (c) Validation of test, (d) Scaling the test, (e) Determination of reliability, objectivity, norms, and perhaps standards.

Of these five steps one is usually omitted in whole or in part, and more often *in whole* than *in part*. Unfortunately the omitted step—validation of the test—is the most important step, for though a test be perfect in every other respect, it is worthless and may be actually harmful if it does not measure what it purports to measure.

Test validation has not been omitted because test constructors have failed to realize its importance, nor has it been omitted because test constructors are lazy. A test can be validated only by demonstrating that it yields a substantial positive correlation with some objectively defined criterion outside of the test which is being validated. Test validation has been omitted frequently because of the great difficulty of objectively defining the criterion. It may be possible to evolve a general procedure and it is certainly possible to

formulate specific procedures for the construction of criteria for evaluating not only mental tests, but also other measurements of individuals. Here is a fit task for both genius and industry.

2. *Area, Volume, or Weight of a Mental Function.* The size of a man is not measured by his height, or width, or thickness, but by a certain balanced combination of these aspects of size. A man's volume is the real measure of his size, for volume gives to every dimension of size its proper influence. Because of a reasonable uniformity in the contents of men, weight is a fairly satisfactory combiner of a man's size dimensions.

Mental traits have area or volume very much as a man has size. There was once a time when no method had been evolved for determining the volume of physical objects. Mental measurement is in exactly that stage of evolution at present. Reading ability, for example, has three dimensions at least—accuracy, speed, and difficulty of the passage read. Furthermore, that experimenter is lucky or obtuse who has not been forced to worry about how to weight time, or errors, or other aspects of mental performance.

The common current method of determining the volume of a mental function is a purely impressionistic one. Estimations of the volumes of objects are risky. Subjective judgments of the volumes of mental traits are truly hazardous. Even the most competent judge has little confidence in his estimate.

I suggest the following methods for determining the volume of a mental function, and express the hope that some of us will find the opportunity to determine the correspondence between them, evaluate each, add others, and finally construct a formula for computing the volume of the mental functions measured by the more widely used standard tests.

(a) Subjective estimate. Has the pupil who reads with a speed of 20, an accuracy of 20, and a difficulty of 20 more reading ability than one who reads with a speed of 25, an accuracy of 15, and a difficulty of 21? (b) Subjective estimate of social worth. It might, for example, be argued that while high speed and low accuracy are as brilliant an illustration of neutral behavior as correspondingly lower speed and greater accuracy, still, the latter should be given a higher score because of the relatively greater social value of accuracy. (c) Weight, i. e., some one-dimensional measurement which represents a condensation of speed, accuracy, and difficulty just as

the weight of a man is a condensation of his size dimensions. (d) Correlation with some criterion which is otherwise defined and objectively measured in some one-dimensional manner. Once the criterion is secured this can be done by means of partial correlation and a regression equation. Once given an intelligence criterion, for example, it is a simple matter to determine the optimum weightings for rights, errors, speed, etc., in a proposed intelligence test. (e) Experimental determination of the equality of various combinations of speed, accuracy and difficulty. Courtis is probably the only one who has made an extensive use of this method. (f) Self-correlation between two trials of equivalent tests upon identical pupils. This method assumes that the proper weighting for each of speed, accuracy, and difficulty in the composite of the three is that weighting which will yield the highest self-correlation between the two composites. There is a statistical procedure which will facilitate the determination of these weightings. The self-correlation method has never been used for determining the volume of a mental trait.

NOTES ON ARTICLES IN EDUCATIONAL PSYCHOLOGY IN CURRENT ISSUES OF OTHER MAGAZINES

Equality in Difficulty of Alternative Intelligence Examinations. Edward L. Thorndike, *Journal of Applied Psychology*, 1920, 4, 283-289, suggests a method of constructing alternative forms of tests and gives data secured with different forms of the Thorndike Intelligence Examination.

Correlation of Army Alpha Intelligence Test, with Academic Grades in High Schools and Military Academies. H. E. Burtt and G. F. Ayrs. *Journal of Applied Psychology*, 1920, 4, 289-294. Finds a correlation of about $+0.40$.

A Comparative Study of the Intelligence of Seventy-five White and Fifty-five Colored College Students by the Stanford Revision of the Binet-Simon Scale. S. M. Derick. *Journal of Applied Psychology*, 1920, 4, 316-330. Whites yield an average I. Q. about 9 points higher than negroes.

The Selection of Mill Workers by Mental Tests. Arthur S. Otis. *Journal of Applied Psychology*, 1920, 4, 339-342. A high degree of intelligence a detriment for certain tasks.

Intelligence Tests and Academic Standing. Edna Gordon and H. J. Baker. *Journal of Applied Psychology*, 1920, 4, 361-363. Correlations range from 0 to $+ .54$, depending on the course of study.

Picture Completion. Robert H. Gault. *Journal of Applied Psychology*, 1920, 4, 310-316. Suggestions concerning the weighting of responses in the Healy test.

Reconstruction of Mental Tests. Beardsley Ruml. *Journal of Philosophy*, 1921, 28, 181-185. A reply to an article by Pressey on statistical methods.

A Comparison of Two Methods of Giving the Number Series Completion Test. John E. Anderson. *Journal of Applied Psychology*, 1920, 4, 346-348.

Education of Juvenile Delinquents. Edgar A. Doll. *Journal of Delinquency*, 1921, 6, 331-347. A discussion of methods of diagnosis and educational treatment.

The Importance of Physical and Mental Examinations as an aid to Treatment and Training in a Reform Institution. Edmund B. Hilliard. *Journal of Delinquency*, 1921, 6, 347-355.

Irregularity in Intelligence Tests of Delinquents. Julia Mathews. *Journal of Delinquency*, 1921, 6, 355-362. Data presented suggests the need of a more careful inquiry into the causes of "scattering."

An English Form Test. Thomas B. Briggs. *Teachers College Record*, 1921, 22, 1-11. Describes two equivalent forms of a test for formal English to be used in grades VII, VIII and IX. Each test consists of 20 short passages to be punctuated by the subject.

Proposed Uniform Method of Scale Construction. William A. McCall. Teachers College Record, 1921, 22, 31-51. Describes a method of scale construction in which the mean performance of children between 12.0 and 13.0 years of age is taken as the reference point, with 10 S.D. as the range and -5 S.D. as the zero point. The range is divided into 100 units, which serve as the scale points. A reading scale constructed by the use of this method is described.

A Qualitative Investigation of the Effect of Mode of Presentation Upon the Process of Learning. F. J. O'Brien. American Journal of Psychology, 1921, 32, 249-284. Suggests the need of supplementary studies of learning by introspective accounts of imagery employed.

The Child Mind. H. J. Mulford. American Journal of Psychology, 1921, 32, 179-196. A variety of opinions on the child's mind from a genetic point of view.

On the Relevancy of Imagery to the Process of Thought. C. Comstock. American Journal of Psychology, 1921, 32, 196-231. A discussion of several functions of imagery in solving problems.

The Use of a Time-Record Blank in the Standardization and Supervision of Student-Teaching Courses. William S. Gray. Educational Administration and Supervision, 1921, 7, 121-133.

The Training of Teachers to Supervise Study. Alfred L. Hall-Quest. Educational Administration and Supervision, 1921, 7, 160-166.

Some Second Lieutenant Psychology in School Administration. Garry C. Myers. Educational Administration and Supervision, 1921, 7, 171-175.

A Measure of Ability to Judge Poetry. Allan Abbott and M. R. Trabue. Teachers College Record, 1921, 22, 101-127. Twenty-six sets of 4 poems each, graded by competent judges, and scaled in steps of merit.

NEW PUBLICATIONS IN EDUCATIONAL PSYCHOLOGY AND RELATED FIELDS OF EDUCATION

1. *Psychology for Normal Schools*.—This text book¹ for normal school students is novel in its organization. Instead of the conventional chapters, the material is arranged in 46 lessons, averaging 7 pages each, introduced by a series of suggestions for observational work and followed by "Topics for Special Study and Report," "The Lesson Applied" and "Selected References."

One hundred and twenty pages are devoted to instinctive behavior, and nearly 50 pages are devoted to the general studies of heredity; then follow 20 pages on habit; 65 pages on sensation, perception and other mental progresses, and a final 60 pages on individual differences. Although the divisions of the book suggest a dynamic psychology, the material is almost wholly descriptive. Instincts are named and described in terms of children's familiar play and other activities. The sections on habit, learning and the mental processes—sensation, perception, attention, imagery, memory, imagination, thinking and will—consist primarily of definitions and illustrations. The laws of learning are nowhere stated; no reference is made to the experimental work upon the acquisition of skill or information, upon problem solving or thinking. No mention is made of a curve of learning or forgetting, of economical methods of memorizing or the psychology of the school subjects; the words spelling, arithmetic and writing do not appear in the index. The chapter on "Memory and Imagination" is typical; memory is defined as "reproductive imagination." Three pages are used to amplify this, and the remainder of the short chapter is devoted to "Earliest memories of childhood" and "Desultory memory of childhood."

The experimental psychology of individual differences is condensed into about 40 pages. Many of these are devoted to definitions of "normal", "subnormal", "epilepsy", "hysteria", etc., but no graph or mention of the curve of distribution is to be found. A half page is

¹*Psychology for Normal Schools*, Lawrence A. Averill, Boston; Houghton Mifflin Company, 1921, Pp. XX + 362.

devoted to the discussion of educational tests; two-thirds of a page to group intelligence tests.

The author, in the preface, finds fault with current texts for normal schools because of "their scientific terminology and style of language," which he has sought to remedy. Whether the cure is better than the malady, the reader may judge from the following sample effort to explain "introspection" as a method of investigation in psychology. The teacher had told a story of primitive firearms which aroused a pupil to exclaim, "My father's got one of them guns up in our shed-chamber." To this the teacher retorted, "How often must I tell you, James, to be careful of your English" The author adds: "Now, you will understand better why introspective cleverness is a virtue on the part of a teacher. James' teacher was not an introspectionist; she failed to see behind James' ungrammatical response the real impulsion that motivated him; she saw not the thing that James saw, but only the poor language in which it was coached."

A. I. G.

2. *Psychology for Nurses*.—The popular tendency to apply the term psychology to any suggestion or opinion concerning the behavior of human beings is exemplified in this book.² A great deal of advice for nurses and many observations of people generally are grouped under the headings of textbooks in psychology. Occasionally a scientific fact or principle is stated and illustrated by hospital experience, but most of the psychology consists of definitions quoted largely from Thorndike's "Elements" and James' "Briefer Course," which appear in an appendix to each chapter. The only other authors mentioned are Pillsbury, Colvin and Bagley, and in one place Terman. The following quotation, picked at random, from the chapter on "Sensation," illustrates the procedure.

V. DERMAL SENSATIONS.

The Nurse's Use of Sensations of the Skin. Sense-organs of the skin have much to do with our responses. The nurse, like the surgeon, needs "eyes in her fingers." The nurse gets all manner of information by way of touch. It

²The *Psychology of Nursing*. Aileen Cleveland Higgins, G. P. Putnam's Sons: New York, 1921. Pp. XVI + 337.

is useful to be sure of one's hands in an emergency, particularly when they must act as guides in darkness or half-light. To gauge accurately the amount of pressure necessary in various nursing procedures is fundamental in good technique.

3. *Psychology for Musicians*.—The first 60 pages of this small book³ give tabulations and discussions of the results of a questionnaire study in which 100 musicians and composers gave replies to a series of 24 questions with regard to methods of learning and teaching, and the effects of various factors on performance. As is usual in such studies, it is difficult to judge what some of the questions and many of the answers mean. It appears that no one type of imagery is employed in learning and performing, although visual, auditory and kinaesthetic representations are most (and about equally) frequently employed. In learning "it is not the sensory field that aids so much as it is the purely intellectual processes." Sixty-six musicians learn "bit by bit," 14 by the "whole method" and 14 by "various combinations of the two." Many assert that emotions (kinds not specified) assist to make performance successful, but add that the emotion must be kept under control. Twenty-one play better when alone, while 65 are favorably stimulated by a sympathetic audience. Nervousness, anger, excitement or disturbances in the audience generally reduce efficiency. Most musicians believe that native aptitude rather than any special form of practice is responsible for their success.

The second part of the book contains a rather superficial exposition of the principles of psychology, consisting chiefly of a series of descriptive definitions of the conventional mental processes. A. I. G.

4. *The Work of a Juvenile Court*.—Franklin C. Hoyt,⁴ Judge of the Children's Court of New York City, has described the principles and machinery upon which the practice of dealing with juvenile offenders is based in a series of narratives designed for the popular reader. The book "does not pretend to cover any particular phase of child psychology, nor is it written with the slightest idea of serving as a manual on juvenile-court work in general." Since it de-

³*The Musician's Mind*. Antoinette Feleky, New York: Pioneer Publishing Company, 1921. Pp. 108.

⁴*Quicksands of Youth*. Franklin Chase Hoyt, New York: Scribners, 1921. Pp. XI + 241.

scribes at length the action of the court in a series of real cases, it will be of interest to many students of education and psychology.

A. I. G.

5. *Two Books Dealing With The Principles of Education.*—Taking as a thesis “the claim of individuality to be regarded as the supreme educational end,” Professor Nunn has written a stimulating survey^a of the whole field of educational theory and practice. The author is throughout alive to scientific research, movements as recent as the use of intelligence tests for admission to Columbia College being considered. The book, in fact, gives an interesting interpretation of current scientific work. In discussing the nature of development, the writings of Pearson, Spearman, Binet, Terman, and especially those of McDougall and Shand, Hart, Freud, Adler, Prince and others are mentioned. The technical studies of Book, Huey, Ach, and Burt are considered. The author marshals experimental findings and current opinion to support his theory that the main function of the school is to socialize its pupils, but points out in a final chapter that such a function “in no wise contradicts the view (stated at the beginning) that its true aim is to cultivate individuality.”

Professor Coursault's book,^b while leading to a somewhat similar conclusion, has a quite different approach. It is based primarily upon theoretical writings in education rather than upon scientific investigations and theories in related fields. The influence of Dewey, Royce, MacVannel, Bagley and McMurry appears clearly. The author attempts to harmonize the views of man as a psycho-physical organism in a process of adjustment through stimuli and responses with the teleological view of man as a person controlled by purposes and ideas. The importance in the educative process of the appreciation of values and the development of purposes as compared to the importance of information is stressed. In explaining how purposes are developed and means of control are made, the author analyzes, in a fashion similar to Dewey, the processes of thinking. In this connection the nature and function of history, literature and

^a*Education, Its Datan and First Principles.* T. Percy Nunn, New York: Longmans, Green and Company, 1920. Pp. VII + 224.

^b*The Principles of Education.* Jesse H. Coursault, Boston: Silver, Burdett and Company, 1920. Pp. XII + 468.

other fine arts and the methods with which subject-matter should be taught are explained. The book will serve as a useful introduction to current educational theory and practice. A. I. G.

6. *A Practical Textbook Interpreting Biological Psychology.**—Rarely do we find a textbook on psychology which is designed to meet the needs of teachers in service and those preparing to teach. If this text had the word "Psychology" in its title, the author might have been accused of using the word, in view of the fact that the contents are not "pure" psychology. Any text which is to lead teachers to make use of the findings of research in the solution of their problems must seek to present its materials in such form that the practical bearings of the principles of psychology are made apparent in illustrative instances. It must also provide some practice in the solution of such problems if instruction is to function in practice.

At the risk of appearing too discursive, this text has interpreted those fundamental conditions governing mental development and shown their relation to the problems of the classroom and the school. In this the author has made his text conform to the psychological requirements of a limited field, avoiding technical and speculative discussions which might make for a more systematic and comprehensive but a less dynamic treatise.

The book is written around two questions: 1. How does the individual normally respond at different periods in his development to the typical situations—physical, intellectual, esthetic and social—in which he is placed? 2. How can he best appropriate the materials and benefits of education so that he can utilize them to greatest advantage in daily life? Part I deals with the dynamic aspects of mental development. Part II restates these and gives educational interpretations. Part III provides exercises in analysis, interpretation, investigation and application. This problem material is full of quotations from other writers, all carefully correlated with the preceding chapters of the book. It occupies over one hundred pages, and furnishes a wealth of opportunities for practice in the sort of thinking which is reflected in improved teaching. L. Z.

*M. V. O'Shea. *Mental Development and Education*. The MacMillan Co., New York, 1921. VII 403.

7. *A Study in the Correction of Speech Defects.*¹—This pamphlet reports the analytical study of the speech disorders of over one hundred children, with a classification of defects and their causes. Two type cases are reported in great detail, with accompanying graphic records of the effect of remedial training and a controlled regime. These case studies are perhaps the most valuable and illuminating part of the contribution. They are followed by materials for speech examination and individual speech records, which make it possible to measure improvement objectively. By making pupils aware of their progress in overcoming defects such records may become a psychologically potent remedial factor. Other remedial suggestions are given, together with a bibliography. L. Z.

8. *A Report on Teacher Training Departments in High Schools.*² Minnesota has attempted to improve the quality of instruction in her rural schools by offering normal training courses in the high schools of small towns. The movement began about twenty-five years ago in a very small way, and was brought to the efficiency described in the report, under the direction of Miss Mabel Carney, whose program included special college or normal training for the instructors, uniform and definite courses throughout the State, and practice teaching in rural schools. The movement received an impetus when increased State aid made more balanced courses possible, and when the completion of the prescribed courses resulted in exemption from teachers' examinations. The plan does not pretend to be the ultimate solution of the problem of providing trained teachers for rural schools. It has been a successful temporary expedient. The program toward which the State Department in Minnesota is working contemplates equal training for city and rural teachers and close affiliation with normal schools. In the meantime, the training departments in high schools are being improved to meet the immediate needs more satisfactorily. The appendix contains (1) a comparative table showing facts concerning training departments in 13 other States, (2) a suggestive yearly scheme for practice teaching, (3) reports, (4) a bibliography. L. Z.

¹Sarah M. Stinchfield. *A Preliminary Study in Corrective Speech*. Studies in Child Welfare, Vol. I, No. 3. University of Iowa, 1920. Pp. 36.

²Lotus D. Coffman. *Teacher Training Departments in Minnesota High Schools*. General Education Board. New York, 1920. Pp. VIII + 92.

9. *Report of the Proceedings of Schoolmen's Week.*⁹—Because the contents of this yearbook duplicate the variety and scope of the program of Schoolmen's Week, it is impossible to give any adequate idea of the material in this limited space. While there is a preponderance of administrative problems, there are only two discussions dealing definitely with curriculum reconstruction. Journal readers will be especially interested in the report of the Bureau of Educational Measurements and in a number of articles dealing with intelligence tests and their uses.

L. Z.

10. *The Educational and Intellectual Status of the Army Medical Service.*¹⁰—Records for approximately 2500 medical officers were available for the statistical investigation reported in this bulletin. The assembled data justify the statement that "the Medical Corps obtained the services of the ablest as well as the weakest men of the profession." The general intelligence rating of medical officers is lower than that of other branches of the service, with the exception of the Dental and Veterinary Corps, and is practically the same as that of the Quartermasters' Corps. But when the psychographs or curves representing measurements for each of eight types of tests are studied, typical differences between those of the medical corps and other arms of the service noted. Those of the several medical groups are similar in certain respects. The median length of schooling was 15.8 years, and 11.07 was the median in years of experience. There is a remarkable difference in the frequency of superior and very superior intelligence in the Medical Corps as compared with the Engineer Corps, and the difference is distinctly in favor of the Engineers, as is shown by tables and graphs. The most significant tabulations and graphs are those in which the achievement in each of the eight parts of the intelligence tests are set down for the various arms of the service, and the various branches of the medical service.

Although the admission of men of inferior ability was due largely to the emergency, it remains true that these men would have practiced medicine even though barred from the army by more rigid re-

⁹*Seventh Annual Schoolmen's Week Proceedings*: University of Pennsylvania Bulletin. Vol. XXI, No. 1. Philadelphia, 1920. Pp. 336

¹⁰M. V. Cobb and R. M. Yerkes. *Intellectual and Educational Status of the Medical Profession as Represented in the United States Army*. Bulletin of the National Research Council. Vol. I, Part 8, No. 8. Washington, D. C. 1921. Pp. 75.

quirements. This study inclines one to the belief that intelligence requirements should be set up for admission to the medical schools and to the profession. A calling so closely related to the national well-being needs the best minds as well as the best training. L. Z.

III. ADDITIONAL PUBLICATIONS RECEIVED.

A. BOOKS IN GENERAL AND APPLIED PSYCHOLOGY.

PATON, STEWART. *Human Behavior*. New York: Charles Scribner's Sons, 1921. Pp. V + 465.

TRACY, F. *Psychology of Adolescence*. New York: MacMillan Company, 1921. Pp. X + 246.

B. MENTAL AND EDUCATIONAL TESTS.

COURTIS, S. A., and SHAW, L. A. *Courtis Standard Practice Tests in Hand-writing*. New York: World Book Co., 1921.

CHAPMAN, J. CROSBY. *Trade Tests*. New York: Henry Holt & Co., 1921. Pp. IX + 435.

TEABUE, M. R., and STOCKBRIDGE, F. P. *Measure Your Mind*. Doubleday, Page & Co., 1921. Pp. VII + 349.

C. PUBLICATIONS IN THE GENERAL EDUCATIONAL FIELD.

FINNEY, R. L. *The American Public School*. New York: MacMillan Company, 1921. Pp. XIV + 335.

O'BRIEN, J. A. *Silent Reading*. New York: MacMillan Company, 1921. Pp. XVII + 289.

SLEIGHT, W. G. *The Organization and Curricula of Schools*. VIII + 264.

D. NEW SCHOOL TEXTBOOKS.

AMES, E. W., and ELDRED, A. *Community Civics*. New York: MacMillan Company, 1921. Pp. XIV + 387.

BURCH, H. R. *American Economic Life*. New York: MacMillan Company, 1921. Pp. XI + 533.

HATFIELD, W. W. *Business English Projects*. New York: MacMillan Company, 1921. Pp. XV + 303.

WASHBURNE, C. W. *Common Science*. New York: World Book Co., 1920. Pp. XV + 390.

The Multum in Parvo Atlas of the World. Chicago: A. J. Nystrom & Co., 1921. Pp. 46.

E. PUBLICATIONS OF UNITED STATES BUREAU OF EDUCATION.

Organization of State Departments of Education. Bureau of Education. Bulletin No. 46, 1920. Pp. 48.

Education for Highway Engineering and Highway Transport. Bureau of Education. Bulletin No. 42, 1920. Pp. 134.

Statistics of State Universities and State Colleges. Bureau of Education. Bulletin No. 87, 1919. Pp. 28.

Statistics of City School Systems, 1917-18. Bureau of Education. Bulletin No. 24. Pp. 477.

F. MISCELLANEOUS PUBLICATIONS.

COLLEGE TEACHERS OF EDUCATION. *Studies in Education.* Educational Monographs No. X. Maryland: King Bros., 1921. Pp. 79.

POWERS, S. R. *A History of the Teaching of Chemistry in the Secondary Schools of the United States Previous to 1850.* Minneapolis: University of Minnesota, 1920. Pp. 68.

The Preliminary Report of the Association Commission on the Organization of the College Curriculum. Chicago: Association of American Colleges Bulletin, 1921. Pp. 60.

ARNETT, T. *Teachers' Salaries in Certain Endowed Colleges and Universities in the United States.* New York: General Education Board, 1921. Pp. 42.

THE CARNEGIE FOUNDATION FOR THE ADVANCEMENT OF TEACHING. *Fifteenth Annual Report of the President and of the Treasurer.* New York, 1920. Pp. 171.

THE JOURNAL OF EDUCATIONAL PSYCHOLOGY

Vol. XII

SEPTEMBER, 1921

No. 6

AN EXPERIMENTAL AND STATISTICAL STUDY OF READING AND READING TESTS.¹

ARTHUR I. GATES

Teachers' College, Columbia University.

A recent bibliography² contains titles of 18 tests for silent reading, 3 for oral reading and 6 for word knowledge (vocabulary). This list bears witness to a keen interest and productive work in the measurement of reading ability, and it is encouraging that in the opinions of many, some of the latest tests are better instruments than most of the earlier ones. While it is doubtless desirable that the abilities of some should be devoted to the creation of new and better tests, it is imperative that extensive experimental and statistical studies be made of the many tests now available, if we are to make rapid progress in the improvement of testing materials.

Mrs. May Ayres Burgess has set an admirable example by accompanying her recent test with a monograph³ in which she has stated carefully the principles upon which her work is based together with an account of the construction of the scale, grade norms, measures of reliability and other necessary information. Too frequently scales are published before their usefulness is known. If the author of a test does not empirically discover its merits and defects, most

¹This study was made possible by the generosity of Mr. Frank A. Vanderlip and the interest and co-operation of Mr. Wilford M. Aiken, Founder and Director, respectively, of the Scarborough School at Scarborough, N. Y. At the beginning of the Academic year 1920-21, a Department of Educational Research was organized in the School, under the direction of the writer, and during the year he has enjoyed the able co-operation of Miss Jessie DeSalle, who was primarily responsible for the testing, and Miss Ella Woodyard, primarily responsible for the statistical work. The co-operation of the teaching staff has been excellent.

²*Bibliography of Tests for Use in Schools.* The World Book Co., Yonkers, N. Y. 1921.

³May Ayres Burgess, *The Measurement of Silent Reading.* Russell Sage Foundation, New York. 1921. Pp. 163.

frequently it is not done at all. The fact is that at the present time we have practically no information concerning most of the tests, outside of tables of norms and possibly a few measures of consistency as determined by retest. We do not know whether they measure rate of reading, comprehension, both or neither. We do not know how speed is related to comprehension, as a consequence. We do not know whether different tests of "comprehension," for example, measure the same or very different functions. We do not agree as to what should constitute a criterion of reading ability. In fact, two writers⁴ have recently questioned the very possibility of testing "general reading ability" by a single scale. "The usual silent reading scale may be considered to measure—not silent reading ability in general, since there seems to be little evidence of any general factor of outstanding importance." P. 29. While the evidence presented by these writers in support of their opinion is very meagre (inter-correlations of 4 short tests given to one grade), it betrays a startling dearth of evidence to the contrary.

There is pretty fair agreement that reading ability depends upon at least two elements—speed and comprehension—but just what shall constitute a measure of either is a matter of dispute. What competent workers think constitutes comprehension, for example, may be discovered by examination of existing tests, of which representatives are listed on a later page. Which of these, if any, best represents general ability to comprehend in reading, we do not know. Some hold that those who utilize an unassisted reproduction, e. g. Brown, are not testing comprehension in reading, but memory, ability to write English compositions and other abilities. Some hold that the scales which present brief paragraphs followed by questions, or directions to cross out a word, etc., measure the ability to reason, to infer, to solve puzzles, to resist suggestion, to attend closely, to discriminate between words, etc., none of which can be called precisely ability to read, although it is possible that all these may be involved in it.

In the interpretation of our results, the assumption has been made that general reading ability is not a fiction, but probably a very broad function. Like general intelligence, it is a reality, but not a single capacity, function or power. It is precisely a cross-section,

⁴L. W. Pressey and S. L. Pressey. *A Critical Study of the Concept of Silent Reading*. *Journal of Educational Psychology*. 1921. 12, 25-32.

average or composite of many functions. What functions are to be included in the composite are in the first instance determined by competent judges. Experimental work results in eliminations and additions. In the case of reading, the assumption has been made that a composite score made up of a number of representative tests, carefully given, does represent general reading ability, and any test no matter what it may appear to be, is a test of reading ability if it yields a satisfactory correlation with this criterion. It is, of course, to be understood that, after experimental work has been done, a much more adequate criterion can doubtless be constructed.

For the purpose of evaluating the usefulness of the several instruments other criteria are to be employed. They are, in the main:

1. Reliability or consistency. Do the subjects perform identically on each of several occasions?
2. Objectivity. Will different experimenters secure the same results?
3. Are the tests units properly equalized or defined?
4. Are the standards (norms) of achievement satisfactory?
5. Does the test properly differentiate or differentiate with satisfactory fineness, or register a sufficiently wide range of abilities?
6. Are the various editions (forms) of the test equivalent?

There are other criteria of more or less practical importance, such as cost, convenience to give or score, time required and interest developed among those taking it.

THE EXPERIMENT IN GENERAL.

This investigation was conducted during the past year at the Scarborough School, Scarborough, New York. It was part of a more extensive study of the constitution of reading ability, with special reference to reading disabilities. All told, about a dozen reading and vocabulary tests were used, along with a greater number of tests of more specific abilities used for purposes of diagnosis. The present report is limited to the material obtained from a group of representative reading and vocabulary tests.

The Subjects: The pupils of Grades III to VIII, inclusive, in the Scarborough School served as subjects. Each grade includes ap-

proximately 20 pupils. The records were complete for each pupil, since absentees were given the tests on reappearance at school.

With very few exceptions, the pupils of the Scarborough school are above the median intelligence of the general population, according to Terman's norms. The median Stanford-Binet Intelligence Quotient for the grades considered is approximately 116.0. This restriction of the range of intelligence should be kept in mind in interpreting the correlations.

The Tests Used: 1. Brown's Silent Reading Test, Forms I and II.

2. The Burgess Scale, P. S., No. 1, given twice.
3. Courtis' Silent Reading Test, No. 2, Forms I and II.
4. Monroe's Silent Reading Test.
5. Thorndike's Scale for the Understanding of Sentences, Alpha 2.
6. Thorndike-McCall Reading Test, Forms 1 and 2 in all grades; Forms 1, 2, 3, 4, 5 in grades IV and VI.
7. Gray's Oral Reading Test.
8. Woodworth-Wells Directions Test.
9. Holley's Sentence Vocabulary Test.
10. A Vocabulary Test arranged by the writer.
11. A Pronunciation Test arranged by the writer.
12. Thorndike's Visual Vocabulary Test. Used in all grades, but, due to an error in administration, results of but two grades were reliable.

The Composite Ratings for Speed and Comprehension. A Composite for Speed: The "rate" scores of the Courtis, Brown, Monroe and the Burgess, each weighed roughly as the square root of the time taken.

B. Composite for Comprehension Number 1: The "Comprehension" scores of the Brown, Courtis, Monroe, Thorndike-McCall, and the Directions test, each weighed roughly as the square root of the time taken. This composite was used before any information concerning the correlations with the individual measures were available. When it turned out that the Brown measure of comprehen-

sion gave approximately a zero correlation with the composite, a new one was constructed. This is called the C. Corrected composite of comprehension: Same as Number I, except that the Brown score is omitted.

Intelligence Tests. In addition to reading and vocabulary tests, the following measures of intelligence were secured:

1. The Stanford Revision of the Binet Scale were given to Grades 3, 4, 5 and 6.

2. A composite of the following group tests, each weighed roughly according to the time: Dearborn, Parts 1, 2, 3 (Grade III), Parts 4 and 5, Grade IV and up; The National Intelligence Scale, Parts A and B, all grades; Otis Primary, Form A, Grades III and IV, Advanced Form A, Grade V and up; Meyer's Mental Measure, all grades, Haggarty, Delta I, Grade III, Delta 2, Grade IV and up; Illinois, all grades; Holley Sentence Completion, all grades, and Terman's Group Test, Grades VII and VIII.

The Coefficients of Correlation. Coefficients of correlation, each test with every other and with the composites were computed for each grade, the Pearson Product Moment formula being used throughout. Corrections for attenuation and the use of the technique of partial correlations in certain instances would add somewhat to the information secured, but the task of computing more than a thousand correlations presented in this paper was so great that further statistical analysis could not be attempted at this time. As measures of central tendencies and of variability the arithmetic mean and the standard deviation have been used.

In dealing with small grade groups, interpretation from coefficients of correlation must be made with very great care, partly because correlations do not necessarily indicate cause and effect or identity of function, and partly because the degree of correlation is dependent upon the range of performance which the group displays. We cannot pretend to have secured the relations of, say, speed and comprehension that would obtain with ideal materials and groups, but corrections for attenuation and for the restriction of range would certainly make them larger than they appear in this paper. A technique for correction of attenuation is available, but laborious. No technique has been devised for correction of restriction in range

of performance. Grades are select groups, and our children are selected entirely from the upper 50%, mostly from the upper 25% of the population. We do not know what the S. D.'s in our tests from a random selection of the universe of children would be. We could not make exact corrections if we did. Some work with our material and in other studies shows a very high correlation between the r 's and the S. D.'s of the measures. Our best assumption is that if all S. D.'s (other things being equal) were as large as the largest, the correlations would be as large as the largest. For purposes of comparing one test with another, which is really our main concern in this study, the data are adequate. Where the S. D.'s are so exceptional as to considerably affect the r 's it will be noted.

The order of presentation will be (1) a survey of the general results with reference particularly to the validity of the concept of general reading ability, and (2) an intensive study of certain features of the several tests, treated singly for the purpose of discovering more exactly what the scales do test and how well they test it.

The Concept of General Reading Ability. Table I gives the means and standard deviations of the correlations of Grades III to VIII, inclusive, for the several tests and the composite scores of comprehension and rate. The facts are an ample justification of the concept of general reading ability, especially when one recalls that no corrections have been made for attenuation or for the decided restrictions of the range of abilities. When it is realized that the criterion for comprehension in some grades represents as many as 8 hours of reading, under test conditions, of a wide variety of materials; connected stories; short, easy directions; short, hard directions; longer paragraphs of directions, easy and hard; paragraphs for interpretation of varied difficulty on all kinds of content, prose and poetry, and these at different times, the fact that a five-minute test (Burgess) should show a mean correlation with it of .8 speaks well for the usefulness of the concept and the test. With the exception of the Brown test, all measures of comprehension yield correlations of .7 or better. Any single test of rate correlates .6 or better, and the oral reading test or a vocabulary test is about as high.

The correlations with the composite of Rate are strikingly similar, most tests measuring one about as well as the other; the slight differences being largely due to the fact that the "comprehension" scores and the "rate" scores are included in their respective com-

posites. The correlation between these composites (which are independent in content) averages $.84 \pm \text{S. D. .08}$. A coefficient of correlation does not enable us to say why this is so. It does not mean, necessarily, that there is no distinction between the two abilities. It shows merely that for some reason the two, in the mass, tend to go together. It will be found later on, as a result of detailed analysis of individual cases, that there is a real and useful distinction between "ability to comprehend" and "rate of reading," and that in such critical cases different tests yield very different scores, and for that reason, where measurement is conducted for purposes of careful individual diagnosis, a test of both "speed" and "comprehension" is essential.

The results do not justify the conclusion that we have, in reading, a group of functions bound by some general factor. The zero correlations yielded persistently by every grade in the case of Brown's test of comprehension is evidence to the contrary; likewise, the low correlations with Stanford-Binet Mental Age and with other functions, e. g. spelling, not presented in this paper. Anticipating material to be presented later, it may be said that the correlation of reading with Mental Age becomes higher as we ascend the grades. This is not the case with the composite of Group Intelligence tests, which throughout yields a fairly high correlation with reading. It is imperative that the relation of reading to other abilities be discovered, and in our attempts we have found it specially instructive to treat the material by grade groups. Many important facts are lost in the massing of material as displayed in Table I.

It is not possible to decide, from the data of Table I, what tests are to be preferred for purposes of measuring general reading ability and for a specific purpose (power of comprehension) what tests are most adequate. This most useful information can be secured only by studying the coefficients of reliability, the grade correlations with the composites, the intercorrelations among tests, and the use of many tests upon unusual types of readers. In the rough, several tests seem to measure the same thing; that is to say, they appear to, so far as can be discerned from coefficients of correlation, but when applied to the peculiar few, distinct differences in the measures appear. We are interested in tests for diagnostic purposes, and they become a means to that end only when we know exactly what they do measure.

THE BROWN SILENT READING TEST.

In a manual of 57 pages,⁵ Brown defends his choice of material and his methods of measuring speed and comprehension. The material is a rather interesting narrative of about 800 words, written in familiar English. There are three forms. The pupil reads silently for one minute, encircling the word last read. The words read per second are obtained by count, giving the conventional score for rate. The pupils are warned that they will be asked at the end of reading to write as much as they can remember. No time limit is set for the reproduction. For scoring the papers, keys are provided, indicating the essential idea in italics and less important matter in plain type. The papers are first read to secure a measure of the "quantity of reproduction," which is stated as the percentage which the amount recalled is of the amount read. The papers are then examined again, "and only those ideas counted which are entirely correct in every respect and of which every detail is reproduced." This is called quality of reproduction and is stated as a percentage. The final comprehension score is the mean of the two. The labor involving in the scoring is surprisingly great and the keys supplied by Brown are not without defects.

On *a priori* grounds, objections could be raised to the method used by Brown, Curtis and others of measured speed by a test which provides no mechanical control of comprehension. In the Curtis test there is no certainty that children are maintaining a uniformity of care as regards comprehension; in fact, there is no certainty that they are comprehending at all. There is the possibility that each child may adopt quite different degrees of care at different times. Brown probably secures more uniformity by virtue of the instruction that the pupils will be asked to reproduce what they have read.

The correlations between two rate tests with the Brown materials are:

⁵Brown, H. A. *The Measurement of Ability to Read*. Department of Public Instruction, Bureau of Research, Bulletin No. 1, 1916, Concord, N. H.

	Coefficient of Correlation	Coefficient of Reliability
Grade III.....	.76	.86
IV.....	.61	.76
V.....	.67	.80
VI.....	.48	.65
VII.....	.57	.72
VIII.....	.60	.75
Mean.....	.61	.76
S. D.....	.09	.09

The coefficient of Reliability⁶ in this case gives the degree, approximately, that the combined results of two trials of the test would correlate with the composite of two other trials. It gives a notion of how consistent the performance of children is. In this test the performances display but a moderately satisfactory degree of consistency, but a degree somewhat higher than that found for the Courtis test. These measures do not tell us anything about the validity of performance. It tells us only that, whether the child reads at a rate consistent with understanding, or whether he skims, or reads with greatest care, he does it with a certain degree of consistency. The validity of the test, i. e. whether it yields a measure of real reading ability, can be discovered by a study of Table II, which contains the correlations with other criteria.

From this table it appears that the Brown Rate Score agrees about as well with other measures of rate as it does with itself. The correlations range downward from .67 with the Monroe Rate, through Directions and Burgess to .53 with Courtis. The correlation with the Composite of Rate is $.82 \pm \text{S. D. .10}$, which is higher than the Courtis, but not as high as the Monroe Rate. The Correlation of the Brown Rate with the corrected composite of comprehension is $.66 \pm \text{S. D. .10}$, which is, again, higher than Courtis, but not as high as Monroe Rate. The correlations with the vocabulary tests are a little better than .4, and with Gray's Oral a little better than .5. The Correlation with Stanford-Binet is very low, averaging $.17 \pm \text{S. D. .12}$, but higher with the group tests of intelligence, $.40 \pm \text{S. D. .21}$.

⁶See Brown, W. *The Essentials of Mental Measurement*. London. 1991. Pp. 101-2. Reliability, two trials, = $2r$

$$1 - r.$$

TABLE II.
Showing the Correlation of Brown's Rate Score with

Grade.	Stanford Binet	M. A.	Comp. Group Tests.	Courts Rate.	Courts Comp.	Brown Comp.	Monroe Rate.	Monroe Comp.	Burgess	Thorndike- McCall.	Direction.	Special Vocab.	Holley Vocab.	Gray's Oral.	Comp. Comp.	Correct Comp.	Comp. Rate.
3.....	17	24	24	87	74	-10	76	67	68	62	70	53	54	57	62	60	83
4.....	03	49	49	43	47	09	74	63	60	24	68	48	45	54	61	65	83
5.....	21	43	43	22	57	07	72	68	75	46	72	60	41	61	83	80	90
6.....	36	52	51	51	49	-02	48	60	41	34	56	32	31	47	60	62	61
7.....	..	69	48	48	60	18	74	30	51	..	65	..	44	50	86
8.....	..	02	65	65	53	-16	50	17	18	..	12	80	91
Mean17	40	53	57	.01	.67	.64	.61	.36	.63	.48	.42	.55	.54	.54	.66	.82
S. D.12	21	.20	.09	.12	.11	.03	.12	.14	.08	.10	.15	.05	.21	.21	.10	.10
Correlation of Brown's Comprehension Score with																	
3.....	40	-08	-09	.08	-10*	-33	-25	07	02	-05	-08	-03	07	.26	10	-19	19
4.....	20	16	-03	-17	09	.28	23	01	37	12	08	05	03	.55	40	54	54
5.....	10	07	-60	06	07	-34	-26	-13	07	-23	-33	-34	-07	.06	-11	28	28
6.....	24	23	-12	-05	-02	.07	-02	-04	25	-01	04	16	06	.13	20	-08	08
7.....	..	09	-02	15	18	16	-24	28	13	.46	10	-02	02
8.....	..	45	05	-20	-16	02	50	24	.52	26	-05	05
Mean	-.02	.15	-.14	-.02	.01	-.08	-.075	.015	.165	.02	-.07	.035	.02	.33	.16	.08	.08
S. D.26	.16	.21	.13	.12	.27	.20	.09	.24	.16	.16	.22	.04	.19	.16	.25	.25

* Brown Rate.

The Brown comprehension score affords the single case of persistent zero correlation with the various measures used. Its correlation with the Brown rate is zero, as it is, approximately, with each and every measure of rate, comprehension, vocabulary and intelligence. Whatever this score does represent, it certainly is not comprehensive in reading unless all other measures are invalid, which is scarcely likely. It may well be that a written reproduction of what was read during a minute is a useful exercise, and needs cultivation, but it is not a measure of comprehension, at least when scored by Brown's method.

Brown suggests a composite score for general reading ability obtained by multiplying the score for rate by the score for comprehension. Such a score has not been used in this study for obvious reasons.

The presence of Brown's score in the composite of comprehension reduces its validity. It was eliminated, consequently, and the corrected criterion used throughout. The data for the uncorrected criterion are printed for whatever statistical interest they may possess.

(To be continued.)

CONSTANCY OF THE STANFORD-BINET I. Q. AS SHOWN BY RETESTS.

HAROLD RUGG AND CECILE COLLOTON
The Lincoln School of Teachers' College

If the Stanford-Binet Intelligence Test is taken two or more times by the same pupils, how closely will the I. Q.'s agree?

At least six reports are now available* from which the answer to this question can be formulated: (a) Terman (see Bibliography, No. 1); (b) Cuneo and Terman (2); (c) Garrison (3); (d) Poull (4); (e) Wallin (5) (d and e appear in this issue of the *Journal of Educational Psychology*; (f) Fermon (6); (g) Stenquist (7).

This article will summarize and interpret the evidence reported by these workers and add evidence secured in the educational psychology laboratory of the Lincoln School of Teachers College, 1920-1921. The data of the six investigations are summarized in Tables I and II. We have incorporated our own data in these tables on bases, so far as possible, which are comparable with those of other studies. The Binet testing in the Lincoln School was done as follows: Of the 137 retests, Mr. Rugg gave 73 initial tests and Miss Anne Brown 64 tests in the winter and spring of 1920. Miss Colloton gave 45 initial tests in 1920-1921 and 121 retests. Mr. Rugg gave 16 retests. Our individual average differences are as follows:

		Number of Retests
Mr. Rugg with himself	5.5**	16
Miss Colloton with Mr. Rugg	4.9**	59
Miss Colloton with Miss Brown	4.5	62

Constancy of the I. Q. can be expressed in three ways: (1) by the average difference between the initial and successive tests; (2) by the limits of the middle 50 per cent of the differences; (3) by the coefficient of correlation between the successive tests. Table I presents these facts for the seven studies. In these studies 1,487 retests are reported. All studies are recent—five, the work of the past year:

*As shown by a search of the following magazines for the years 1915, 1916, 1917, 1918, 1919, 1920, 1921: JOURNAL OF EDUCATIONAL PSYCHOLOGY; *Journal of Educational Research*; *Journal Experimental Psychology*; *School and Society*; *Training School Bulletin*; *Psychological Clinic*; *Psychological Review*, and *Psychological Index*. We will appreciate information from any reader who knows of other published or unpublished studies of Stanford-Binet Retests.

**These average differences become 4.9 and 4.4 respectively if cases are omitted in which the pupils' mental ability was not completely explored at the initial test. This was caused by a rigorous following of directions.

TABLE 1 SUMMARY OF INVESTIGATIONS ON RETESTS WITH THE STANFORD-BINET

Investigator	Date	Number Of Children					Interval	Central Tendency of Change Between 1st & 2nd Tests	Limits of Middle 50%	Average Difference	Coefficient of Correlation
		Total	5-11 Yrs.	11-12 Yrs.	12-13 Yrs.	13-14 Yrs.					
Terman	1918	4315	99	139	134	63	Less Than 1 Yr. - 86 1 Yr. to 3 Yrs. - 138 3 to 5 Yrs. - 85 More Than 5 Yrs. - 127	+1.78	+5.7 -3.3	4.5 (P.E.)	.93
Cuneo And Terman	1918	77	70	7	0	0	2 Days - 25 5-7 Mos. - 21 20-24 Mos. - 31	Median Change = 6	+8 -3		.95 .92 .92
* Garrison	1921	62	0	12	49	1	3 Yrs			4.66	
Poull	1921	126	(4-28 Yrs.)				6 Mos.-3 Yrs.	Average Change \$1.28	+4.8 -3.3	4.6	
Wallin	1921	Scale, 1908-61						Ave. Imp. 6.6 Ave. Reduct. 7.8		6.6	
		1911-61						4.3	7.8	6.2	
		1911-19						8.1	5.1	6.1	
		1911-120								10.2	
		1908-120								14.1	
		1908-120									
Egge - Colleton	1921	137	0	51	50	36	10 Mos. to 1 Yr 4 Mos.	Median Difference \$1.6	+5.6 -2.3	4.7	.84
Fermon (unpublished)	1920	**	233				7 Mos. to 4 Yrs		+15 0	6 Yrs. = 10.0 7 Yrs. = 9.6 7.5 (P.E.)	
Stenquist (unpublished)	1920	**	274	26	198	48	0	Less Than 1 Yr. - 32 1 Yr.-3 Yrs. = 24	Med. Gains 3.5 Med. Losses 2.6		72
* First Test - Goddard Revision Second - Stanford											
** 274 Cases Reported By Stenquist Include The Testing With Pupils In A New York City											
233 Reported By Fermon - Less Fermon Did Not											

The present answer to the question concerning constancy of I. Q. The findings of Fermon and Stenquist are sharply distinguished from those of all other workers. Terman, Terman and Cuneo, Garrison, Poull, Wallin and the present writers report average differences in I. Q. between first and second tests of approximately 5 points I. Q. The investigations by Terman, Garrison and Poull, together with ours, represent 760 children. The average difference for these studies is closely 4.5 points I. Q. This means that the chances are approximately 20 to 1 that the I. Q. of a pupil reported from a single test (as measured in the Stanford-Binet with the care represented by these studies) is within 13 points of his true I. Q.

Middle fifty per cent. For all studies the positive differences are nearly twice as large as the negative differences. Even so, the studies show that typical positive differences are less than 6 points. Typical negative differences are approximately 3 points. This means that the chances are one in two that an I. Q. from a single test will increase as much as 6 points or decrease as much as 3; that the chances are 1 in 5 that it will increase as much as 12, or decrease

decrease as much as 6; that the chances are 1 in 20 that it will increase as much as 18 or decrease as much as 9.

A significant fact therefore: much confidence can be put on a single I. Q. if the examination is made by experienced and well-trained examiners who use rigorously the standardized procedure for giving the test. In a range of intelligence for large bodies of public school children of, say, 50 points (from 80 to 130 I. Q.) it is very helpful to be able to predict intelligence with as much precision as is implied by these figures. Furthermore, the giving of a retest in all doubtful cases will increase the stated degree of reliability by about 40 per cent. That is, for two tests the P. E. becomes approximately 3 points.

Thus, the recent studies, except those of Stenquist and Fermon, closely confirm Terman in his earlier statements.

We have studied the details of the reports by Fermon and Stenquist. The latter are careful to state that the examiners who did the testing were carefully trained and had tested at least 20 pupils under critical supervision. The comparison of their findings with those of the other studies throws great doubt on the validity of the examining which was done by these workers. We are convinced that the great differences in I. Q. must have been caused primarily by non-uniform scoring of responses by those who gave the tests. Stenquist says, however, "it seems certain that the differing I. Q.'s obtained from the successive tests cannot be accounted for by the personal equation of examiners. They are probably due, on the one hand, to actual differences in the child from time to time, and on the other hand to the fallibility of the crude instruments with which we are measuring a most complex thing." (He is careful to state that his criticism is of the Binet scale and the I. Q. as *absolute measures of intelligence*.)

Study the charts presented as Table II. These present a very interesting and important comparison of the detailed distribution of differences in I. Q. The extreme differences in retest are important as well as the central tendencies. It is significant that four different groups of investigators, working independently, obtain differences in retest of more than 10 points in less than one-sixth of the cases. In our own work no difference was greater than +17 or -15; 12% were more than 10. In Terman's 67 out of 435, or 15%,

TABLE II-1 DISTRIBUTION OF CHANGES IN I.Q.
BETWEEN THE 1ST and 2ND TESTS

TERMAN - 435 CASES.

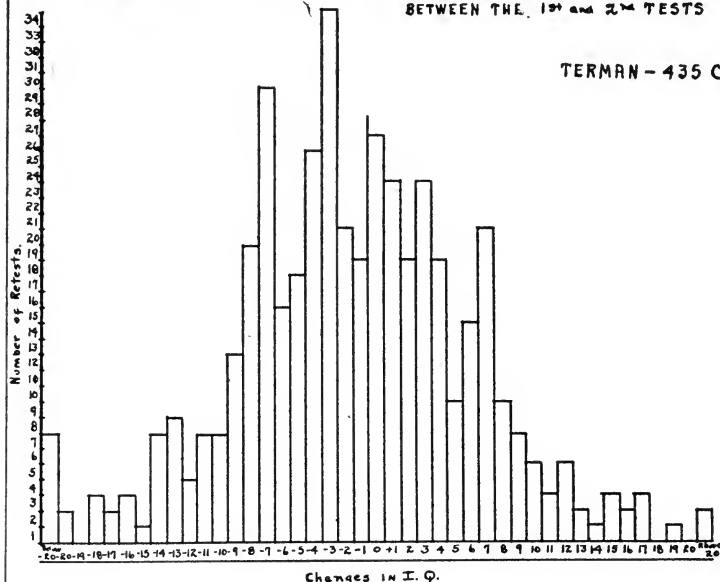
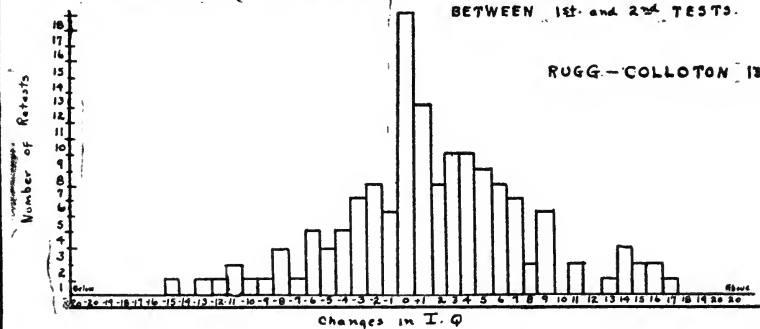


TABLE II-2 DISTRIBUTION OF CHANGES IN I. Q.
BETWEEN 1ST and 2ND TESTS.

RUGG-COLLOTON 131 CASES



GARRISON - 62 CASES

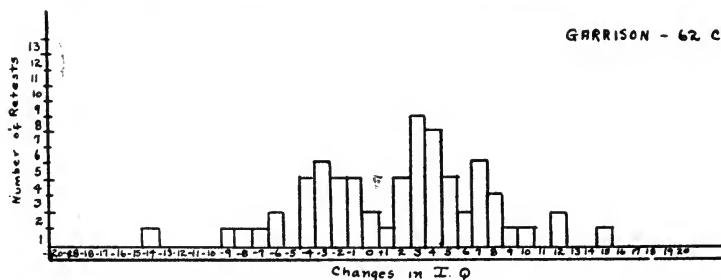
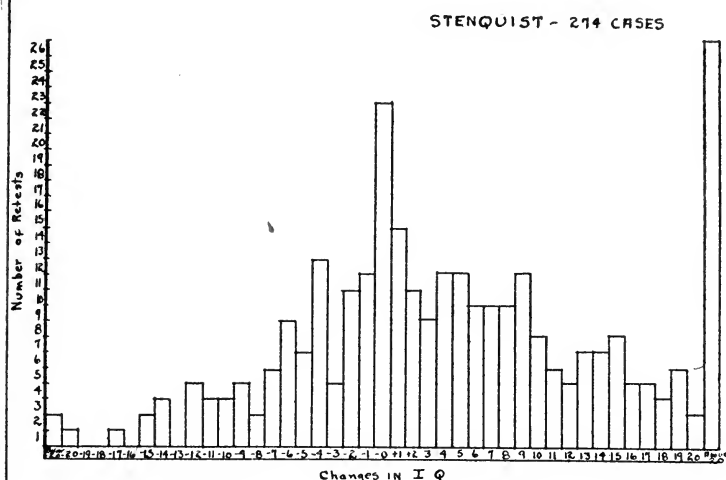


TABLE II-3. DISTRIBUTION OF CHANGES IN I. Q.
BETWEEN THE 1st and 2nd TESTS



were greater than 10 points. Of our 137 retests 23 were greater than 8. *Eight of these can be definitely explained by the fact that the first test did not completely explore the pupil's mental ability. This raises an important point of technique,—that of not carrying the testing far enough to completely explore the pupil's general mental ability.*

Only 6% of Garrison's cases showed differences greater than 10.

Average differences classified according to age of pupils.

Table III classifies the differences by age levels. It shows that these differences are only slightly larger with very young children, especially below the entering school age of 6. It also shows that with no school children do the average differences exceed 7 points.

(These conclusions ignore the data of Stenquist and Fermon, which, as indicated above, must be unsound.) For children of school age our data show that differences are *not* appreciably larger with the younger children, say, 6-9 years. In fact, the difference in difference in retest may be neglected. This is contrary to the common view of the matter.

TABLE III.
Comparison of Average Differences Between 1st and 2nd Tests, Classified
According to Age Levels.

Investigator	3 yrs.— 5 yrs. 11 mo.	Average Difference	6 yrs.— 8 yrs. 11 mo.	Average Difference	9 yrs.— 11 yrs. 11 mo.	Average Difference	12 yrs. and over	Average Difference
Terman	99	6.9	139	6.0	134	5.3	63	6.3
Garrison	0	—	12	3.6	49	4.7	1	—
Stenquist	28	13.5	198	7.7	48	6.9	0	—
Rugg—Colloton	0	—	51	4.5	50	5.5	36	3.7

TABLE IV.
Comparison of Average Differences Between 1st and 2nd Tests, Classified
According to Degree of Intelligence.

Investigator	Bright Above 110 I. Q.	Average Difference	Average 90—109 I. Q.	Average Difference	Dull Below 90 I. Q.	Average Difference
Terman	183	5.8	147	6.2	104	5.8
Garrison	26	5.6	31	4.0	5	9.2
Stenquist	118	8.4	101	8.0	55	8.2
Rugg—Colloton	97	4.6	39	4.7	1	—

Average differences classified according to degree of intelligence of the pupils. Table IV gives the data. The conclusion is the same as in the case of average differences classified on age levels: difference in degree of intelligence seems not to be a factor. Differences in retest will be approximately the same, irrespective of the intelligence of the pupils.

In Table V we present the details of our retests thrown together in one correlation table. We fixed a correlation of .84 between the

TABLE V.
Agreement Between 1st and 2nd Test Rugg-Colloton 137 Cases Correlation
 $r = .84$.

		I.Q. AT. 2 ND TEST																	
		80	85	90	95	100	105	110	115	120	125	130	135	140	145	150	155	160	165
I.Q. AT. 1 ST TEST.	165																		1
	160																		
	155													1	1	1			
	150														1	1			
	145																1		
	140												4	3	1				
	135									1		1	4						
	130									1	1	3	1	1					
	125							1	1	1	7	1	1	1					
	120							3	2	3	3	1	1						
	115							3	16	5		2							
	110						2	7	5	2	3								
	105					2	3	3	2	2									
	100					4	8	3											
	95				2	2	1	1											
	90			2	1														
	85				1														
	80																		1 3 7

first and second test. Terman reports .93, Cuneo and Terman .95, .94 and .85, respectively.

Thus, the new investigations tend to confirm Terman in his 1917 conclusions and to give us much confidence in the constancy of the I. Q. as measured by the Stanford Revision of the Binet-Simon Scale.*

*Nevertheless there is much to be done in improving the scale and, probably, in making new individual scales. We will present definite criticisms of the scale at a later time.

BIBLIOGRAPHY.

1. Terman, Lewis, M., *The Intelligence of School Children*. Chapter IX, p. 135.
2. Cuneo, Irene and Terman, L. M. *Stanford-Binet Tests of 112 Kindergarten Children and 77 repeated tests*, Pedagogical Seminary, 1918, -25, 414-428.
3. Garrison, S. C., *Fluctuation of Intelligence Quotient*. School and Society, June, 1921.
4. Poull, Louise E., *Constancy of I. Q. in Mental Defectives According to the Stanford Revision of Binet Tests*. Journal Educational Psychology, September, 1921.
5. Wallin, J. E. Wallace, *The Results of Retests by Means of the Binet Scale*. Journal Educational Psychology, September, 1921.
6. Fermon, Marcella L., *Validity of I. Q. as Established by Retests*. M. A. Thesis, Columbia, University, May, 1920.
7. Stenquist, John L., *Unreliability of Individual and Group Intelligence Tests in Grades 1, 2, and 3*. (Unpublished: includes data of Fermon, 6.)

CONSTANCY OF I. Q. IN MENTAL DEFECTIVES, ACCORDING TO THE STANFORD-REVISION OF BINET TESTS.

LOUISE E. POULL.

Psychologist, Children's Hospital, Randall's Island, New York City.

The data of Table I were derived from retests of 126 inmates of Children's Hospital, Randall's Island. The retests were made as part of the routine work of the institution and the cases are, therefore, unselected, excepting that epileptics were excluded. The intervals between the tests varied from six months to three years; the ages of the subjects from four years to 28 years; the I. Q.'s of the first test from 20 to 90. The records were made by trained psychologists, accustomed to the reactions of mental defectives.

The data of the table show the plus or minus changes of the second test over the first in points of I.Q. It will be seen, on inspection, that these subjects as a group did not deteriorate. The average change is an increase of $+1.28$. The middle 50 per cent lie between -3.3 and $+4.8$ variation. The Standard Deviation was found to be 5.83 .

It is significant that the curve does not differ from the one worked out by Terman from the records of unselected school children. ("The Intelligence of School Children," p. 141 ff.) The indication is that mental defectives are not more variable than normal subjects. It is not assumed, however, that the question of the constancy of I. Q. is settled until further studies have been made covering regular intervals and including repeated tests of the same cases over a number of years.

The injustice of disposing of persons under suspicion of mental defect on the basis of a single test is clearly demonstrated. A large percentage of the cases shows variations which operate to change the classification, and, in cases above the obvious imbecile type, only observation and re-testing can discover the individuals who require permanent supervision or institutional care. Composite ratings, including non-language and performance scales, together with the Stanford-Binet, have been found to give truer evaluations of problem cases, since they give weight to the manual abilities which sometimes express a degree of intelligence hidden by language inhibitions.

Amount of Difference	Number of Cases
+17	1
+16	0
+15	0
+14	2
+13	2
+12	3
+11	3
+10	4
+ 9	1
+ 8	2
+ 7	7
+ 6	3
+ 5	3
+ 4	6
+ 3	10
+ 2	8
+ 1	11
— 1	13
— 2	5
— 3	9
— 4	10
— 5	4
— 6	5
— 7	3
— 8	3
— 9	1
—10	0
—11	1
—12	1
—13	2
—14	0
—15	0
	1
	124

MENTAL GROWTH AND THE I. Q.

LEWIS M. TERMAN
Stanford University.

The problems relating to native mental abilities and to the developmental changes which come with increasing maturity are, and perhaps always will be, among the central problems of educational psychology. We can not adapt the curriculum to the child without fairly accurate knowledge of the mental abilities which the child for the time being possesses and of the abilities which are required to master given types of curricula. We can not intelligently plan a child's later education without more or less dependable means of forecasting what abilities he will possess at a given time in the future. Anything that adds to our knowledge of mental growth is bound to be of great practical significance for education. Hence the immense popularity of the Binet tests, which for the first time made possible a fairly serviceable determination of the stage of intellectual maturity which a given subject had attained. The value of these tests was even further enhanced when it was discovered that the intelligence quotient maintains, during the growth period at least, a certain amount of constancy. In proportion as laws of intellectual development obtain, the door of the future may be opened; determination of the child's present intelligence status will enable us to forecast, within certain limits of error, what manner of adult he will become.

That rough prediction is now possible on the basis of intelligence tests can no longer be denied. For example, it is a fairly safe prediction that the child who has been competently tested by the Binet scale and found to have an I. Q. of 75 will never attain an I. Q. of 125, or that an I. Q. of 125 will never, barring definite nervous disease, drop to 75. No one would now expect a child with an I. Q. of 60 or 70 to be able to graduate from an average high school or pursue a college course. These predictions are of course very rough, but it is worth something to know that in general there is even a *tendency* for the superior to remain superior, for the average to remain average, and for the inferior to remain inferior. The child-study literature of a decade or so ago gave wide currency to the view that the typical genius was as a child stupid, and that intellectual

precocity is likely to be followed by post-adolescent stupidity! To have progressed in ten years from this stage of ignorance to the point where we can predict with a probable error of 4 or 5 points what I. Q. a given child will have several years hence is a long step forward.

There is no likelihood, however, that even this modest claim for the possibility of prediction will go unchallenged. In fact, insistent challenge has already come, chiefly from two sources. First, from teachers and others whose inclination is to believe in miracles and to look askance at so-called "laws of growth" on the basis of which we presume to forecast a child's future. The acceptance of such laws is hindered by the deep-seated and blind faith that anything is possible for any child. To people who derive satisfaction from the fact that child nature contains so many unknown quantities, the suggestion that one's final intelligence level may be predicted is actually repugnant.

A challenge no less insistent has been voiced by a number of psychologists. It should go without saying that the questions raised by any psychologist who has seriously investigated the problem are entitled to a hearing. The issues are large enough to justify any amount of scientific caution. At the same time it is possible that over-zealous attack on a tentative hypothesis may be as inimical to true progress as its over-zealous and dogmatic support. I think it can be shown, for example, that some of the recent criticisms of the I. Q. are based on arguments and data so questionable that they are less likely to strengthen than to weaken the position they are intended to uphold. In the opinion of some, if the I. Q. can be demonstrated to have less than absolute constancy in a majority of cases, or to be markedly variable in selected individual cases, or to show a decrease with age in the case of feeble-minded subjects, or to be capable of misuse by the ignorant, it is *ipso facto* worthless and dangerous.

It is not the purpose of the present article to defend the I. Q. Whatever merits or faults it may have as an index of an individual's present or future intellectual status will sooner or later be determined by investigation. At present I wish only to examine some of the arguments and data relating to mental growth and the validity of the I. Q.

Let us consider first Dr. Doll's recent monograph on *The Growth of Intelligence*.* This study is based on repeated examinations, over a period of three to five years, of 203 feeble-minded subjects in the Vineland Training School. Of the entire number, 55 had been examined at least once a year for five years and 72 at least every year but one for five years. In the case of 27, re-tests were continued only three years. Of the 203 subjects, 95 were below the age of 15 years at the time of the initial test. Of these, 67 were followed for as much as three years prior to the life age of about 15½ years. The life ages of these 67 at the initial test were distributed as follows:

Life age	6	7	8	9	10	11	12
Number	1	6	15	13	8	13	11

The age range of the 203 subjects was from 6 to 66 years, the range of the initial mental ages from 1 to 10.7, and that of the intelligence quotients from 7 to 88. It should be borne in mind in the following discussion that only the data from these 67 cases can be regarded as significant for mental growth and I. Q. validity. The value of the data from even this small group is greatly impaired by the small number of subjects at each age. The fact that nearly half of the 67 cases were below 50 I. Q. at the time of the initial test means that the study can throw little if any light on the mental growth of normal or merely backward children.

The scale used for the first two years was the Goddard translation of the 1908 Binet scale. For the remainder of the investigation the 1911 Goddard Revision was used. The earlier records were translated, in so far as it was possible to do so, into terms of the Goddard Revision. The tests were given "by a large number of different examiners," a part of them, it appears, by summer school students in training.

Not all of the tests were complete or of sufficiently wide range, but a workable objective method was devised for computing mental age scores in such cases. For each subject the mental growth curve was based upon smoothed data, not upon individual examinations. For example, all the mental ages for a subject in the first two years were averaged and the resulting value was taken as the mental age at the mid-point for this period. Then the mental ages for the second and third years were averaged and the result taken as the mental age

**Psychological Monographs*, Vol. 29, No. 2, Whole No. 131, 1921, pp. 130.

for the second mid-point, and similarly for the third and fourth years, the fourth and fifth, etc. This method has its advantages, but it also has the effect of shortening considerably the final growth curve for each subject and of making it slightly flatter.

For the purpose of establishing average mental growth curves the author classifies his subjects according to the final mental age attained and gives us the average curves separately for the groups with final mental age of 1 year, 2 years, 3 years, etc. The resulting curves are relatively flat, showing that for these feeble-minded subjects there is a marked tendency for the I. Q. derived from the Goddard Revision to decrease. For those whose final mental age is 4 or 5, there is relatively little mental growth after life age 11 and little after 12 for those whose final mental age is 6, 7, or 8. Those who reach the mental age of 9 or 10 continue to develop until 16, according to the data presented, although the author calls it 15. In view of the scantiness of the data for each of the groups, these findings, while extremely interesting, can not be taken as final.

The author discusses at length the question of age at which mental growth normally ceases. As he admits that his morons show improvement up to 15, it is surprising to find him contending that the normal adult level of intelligence is reached at about 13 years. If so, the feeble-minded develop later than the normal, which is not only contrary to the generally accepted view, but also to the frequently reiterated opinion of the author. In fact, the author's treatment of this subject is rather confused and self-contradictory. After presenting his 13½-year hypothesis (p. 9 ff.) he alleges in support of it (p. 13) the argument that my assumption of the 16-year adult level is due to this being the "efficiency limit" of the Stanford Revision, which is of course irrelevant to the question. Then in order as given we find the following statements:

P. 14. [The final arrest] "is probably no higher than 14."

P. 15. "There is reason to believe that the true age of average arrest of mental age growth is actually between 13 and 14."

P. 15. "This age [the age of final arrest for normal subjects] may be 15 or 16 or higher, but for reasons given may be provisionally placed at 13 years."

P. 16. "Observation leads one to believe that idiots are arrested in their intellectual growth very early in life (say at about the life age of 5 or 6 years), that imbeciles are arrested at a somewhat later age (say about 10 or 12 years), and that morons are arrested still later (say about 15)."

P. 59. "The ages for each [final] mental level at which all subjects are arrested are as follows:—

[Final] mental age level	1	2	3	4	5	6	7	8	9	10
Age of arrest	10	8	6	11	12	15	15	12	15	14

P. 68. ". . . The average rate of mental age increase of these feeble-minded subjects . . . reaches a practical minimum at 13 or 14 years."

P. 76. [Morons] "are arrested about 15 years."

P. 84. Apropos of the age of growth cessation in the case of superior children, "presumably the rate of growth would decrease after life age 13 years, since the final mental level of superior children is practically attained at that time."

P. 108. "Nearly all these subjects [feeble-minded] cease to develop *several years before age 16*, the theoretical limit to which they are expected to develop at a constant rate by the I. Q." (*Italics mine.*) The last clause is of course entirely irrelevant, as the validity of the I. Q. does not hinge upon growth ceasing at any particular age.

P. 118. "There is an age of arrest for every feeble-minded subject which almost invariably is reached before 15 years of age"

Even the reader's natural expectation of relief on coming to the author's final summary is premature, for on two pages (127-128) we find the following three conclusions: "The more recent and extensive evidence suggests that the average adult level of intelligence is between 13 and 14." "Significant mental age increases are limited to subjects under 15 years of life age." "It [the annual rate of growth] reaches a minimum at about 13 years of life age."*

However, the author's real adherence is to the 13-year hypothesis, to which he seems to have been led chiefly by the results of army mental testing. One may question whether he has not been too in-

*It may interest the reader to know that the average of the above estimates is 13.83 years, and the mean deviation .69 years.

clined to accept these results at their face value. The now famous mental age of 13.4 found for 653 unselected white enlisted men tested by the Stanford-Binet in August, 1918 (or for that matter other army test results), does not, for the following reasons, indicate the life age at which mental growth ceases:

(1) The Stanford-Binet may be somewhat too difficult in the upper ranges. For all anyone knows, the mental age score 13.4 ought to be really 14.4. The error may be greater or less than this.

(2) The 653 enlisted men for whom the average mental age 13.4 was found cannot be taken as representative of the entire draft army. A disproportionate number of them were from the southern and semi-southern states where, according to all the results of army mental testing, average intelligence is lower than in the northern and western states.

(3) Just as this group may not have been representative of the entire draft, the draft army itself was certainly not representative of the male population between the ages of 21 and 31. Of 9,500,000 registrants between these ages 6,973,000 were given exemption or deferred classification. For example, 67,000 agricultural "managers" and 61,000 agricultural "directors" and "comptrollers" were exempted. Those classified as "farmers" in the draft army were in the main farm laborers. This is only a sample of the kind of selection that occurred all along the line. Probably only a small proportion of men in positions of even minor responsibility were drafted. The fraction of 1 per cent. exempted because of mental inferiority was insignificant in comparison with the exemptions of skilled laborers, business men, and professional men. Furthermore, at the time the 653 men were tested (August, 1918) there were 619,000 men in the military or naval service who had not been drafted, or between 15 and 20 per cent of the entire military and naval forces. There is reason to believe that these volunteers included a disproportionate number of college graduates, college students, recent graduates of high schools, high school students, and high-minded youths from the better classes generally. From such facts it is clear that the intellectual cream of the country between the ages 21 and 31 had been skimmed off several times before these 653 drafted men arrived in camp. Nor should it be forgotten that all officers were excluded from this "enlisted men" group.

(4) The conditions under which tests were given in the army were in most cases far from ideal. The men had just reached camp. Doubtless many of them were bewildered or fatigued. Some were suffering from the effects of typhoid and smallpox vaccinations.

(5) The scale used was an abbreviated Stanford-Binet, consisting of four tests in each age group. While this abbreviation yields scores which correlate very highly with scores from the entire scale, I think it can be shown that with adult subjects they tend on the average to run slightly lower.

(6) A large proportion of the tests were given by examiners who had had little training in Binet procedure. My experience leads me to believe that partially trained examiners are more likely to score the tests too rigidly than too leniently.*

My own 16-year estimate may be too high. As it was frankly tentative, I do not feel called upon to defend it. Fifteen years may be nearer the truth. Fourteen may be, but I doubt it. Anybody's estimate at present is of course only guesswork. We may concede Dr. Doll a right to his own guess without admitting his claim to have overthrown the guesses of others.

It will bear repeating that the author presents no data which throw any light on the age at which growth normally ceases, and that even his data for feeble-minded are, as far as this point is concerned, extremely scanty. Of his 203 subjects, only 95 were below 15 at the time of the initial test; of these, only 67 were followed as much as three years prior to reaching age 15½; of the 67, 17 were not followed beyond age 13; and of the remaining 50, a large majority were of idiot or imbecile grade.

Unless these facts are borne in mind the reader will be continually misled in regard to the amount of growth which may be expected of the feeble-minded. For example, the statement (p. 47) that "only 6 subjects, or 3 per cent, have gained as much as two years in four years of life" is seriously misleading. The 3 per cent figure is based on the entire 203 subjects, only 95 of whom were below 15 at initial examination. Of these, we could expect none to develop two years in four who were below 50 I. Q., even if the I. Q. remained constant.

*For some of the above facts, especially those relating to draft statistics, I have drawn upon a memorandum which I addressed to Major Yerkes on January 27, 1919, while the scientific report of the army mental testing was being prepared. The data for the memorandum were secured from official reports to which I do not now have access and cannot from memory locate, but I think there is no doubt about their essential accuracy.

This throws out 42 more, leaving only 53. But of those above 50 I. Q., we could expect none to add two years to the initial mental age who had not at least four years before reaching the age of 15. This throws out 14 additional cases, leaving only 39. However, we must further eliminate all subjects who had a mental age of 8 or more at the initial test, as the author admits that the extreme limit of efficiency of the Goddard Revision is 10 years. This throws out 6 more, leaving 33. On this basis the 3 per cent. now becomes 18 per cent. Mental growth of the amount indicated is six times as likely to be encountered, as the author's statement would suggest.

We have seen that the author's conclusions are not always in harmony with the raw data which he presents. It remains to point out that his original data are also misleading because of the intelligence scale on which they were based. The author raises this question, but while admitting that the Goddard Revision fails to differentiate above the mental age of 10 years, and that it has certain irregularities below this point, says that "none of these arguments seriously affects our results for the feeble-minded." (P. 122.) Also (p. 125), "the worst we could expect from an admittedly imperfect scale would be to find *irregularities* in our average growth curves"; and (p. 126) "from all these considerations we may conclude that the Goddard Scale is valid for our purposes, no matter from what angle it is viewed."

However, partly from previously published data of my own* and partly from a table of equivalents presented by the author (p. 124), I estimate that a subject whose I. Q. by the Stanford Revision remained steadily at 75 from age 8 to 15 years would have about the following Goddard I. Q.'s at the different life ages:

Life age	8	9	10	11	12	13	14	15
Stanford I. Q.	75	75	75	75	75	75	75	75
Goddard I. Q.	85	84	81	78	76	73	71	68

I also estimate that after Goddard mental age 7 a year of growth as measured by the Goddard Revision equals about 1.2 years by the Stanford-Binet. In fact, the three years of Goddard mental age from 7 to 10 are equal to nearly four years by the Stanford-Binet, since at 7 the Goddard Revision is about .9 of a year easier and at

10 slightly harder than the Stanford-Binet.* The result, of course, is an exaggerated flattening of the growth curves approximately as indicated in Figure 1.*†

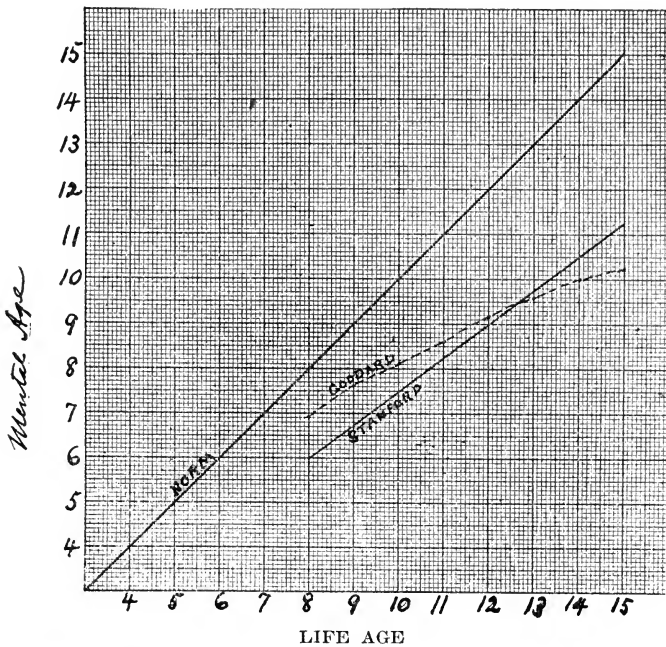


Figure 1. Mental growth curves of the same individuals as shown by the Goddard and Stanford Revisions. (Note: To save space the first three years have been omitted.)

It is therefore impossible to accept the author's statement (p. 77) that "figure 10 [showing average growth curves] lends color to the theory that the upper grades of feeble-mindedness have approximately an average normal rate of growth early in life." That his average mental age curves appear to support this theory is largely due to the fact that those which begin at mental age 5 to 7 years are displaced nearly a year upward. The author departs still farther from his data in the statement (p. 77) that "the average

*Terman and Knollin: Some Problems Relating to the Detection of Borderline Cases of Mental Deficiency. *J. of Psycho-Asthenics*, Vol. 20, 1915, pp. 1-15.

†Thorndike (*Psychological Clinic*, 1914, 8, 185-189) shows that the Goddard mental age norms do not even fit the life ages of the children on whom they were based, being considerably too high in the lower range and considerably too low in the upper range. This is in fair agreement with my own data.

high-grade feeble-minded subject is 'at age' early in life, and is only potentially feeble-minded from the standpoint of mental age." (*Italics mine.*) This is probably true of some feeble-minded subjects, but that it is the rule for any grade of mental deficiency is not indicated by any data known to me. The well-known facts regarding the late walking and late talking of a majority of feeble-minded children suggest that what the author takes to be the rule is decidedly the exception. As for the author's data, only one of his 203 cases was below the age of 7 years at the time of the initial test.

The author devotes 42 of the 130 pages of his monograph to a "Critique of the I. Q." While admitting that "the I. Q. is valuable as a measure of relative brightness" and that it is "superior to mere difference between age and mental age as a measure of retardation" (p. 89), he concludes that it is so lacking in constancy as to be misleading and worthless for purposes of forecasting later mental development or "as a means of classification of such significant intellectual types as feeble-minded or gifted children." He states that "only 1 subject out of a total of 106 feeble-minded subjects who were below 16 years of age at the first examination maintains an I. Q. which is in accord with the theory that the I. Q. is constant." (P. 118.) My own view on the subject, he thinks, is not warranted by the facts.

Notwithstanding his conclusions, the author presents no facts which contradict my own findings. On the contrary, his data for feeble-minded subjects agree much more closely with those I have found for normal children than I should have expected. For the author's 95 cases who were below the age of 15 at the initial test I have computed the correlation between initial I. Q. and that found at the end of three years. I have chosen a three year period for the comparison because a fourth of his subjects below 15 were not re-tested for more than three years. The author does not give the I. Q's. and I have not been able to compute them with perfect accuracy for the reason that the initial ages are given only in whole numbers as 7 years, 8 years, etc. I have therefore treated his 7 year group as though all were $7\frac{1}{2}$, the 8 year group as $8\frac{1}{2}$, etc. The resulting error would presumably be in one direction as often as the other and would not affect the correlation except slightly to

lower it. Even so, the correlation, as shown in Table 1, is .963! While it is true that the I. Q.'s. for these subjects tend to decrease, as is generally admitted to be the case with feeble-minded subjects, this does not interfere with the use of the I. Q. for purposes of prediction. For those who were re-tested for three years before reaching the age of 15½ years the central tendency is toward a drop in that time of 8 points. The upper quartile of changes is at -4.3 and the lower quartile at -11.2.

		I Q at end of three years.																
		5	10	15	20	25	30	35	40	45	50	55	60	65	70	75		
First I Q	85																4	1
	80																	
	75																	
	70																	
	65																1	
	60																2	
	55																4	
	50																1	
	45																	
	40																	
	35																	
	30																	
	25																	
	20																	
	15																	
	10																	
	5-9																	

Table 1. Showing correlation between initial I. Q. and I. Q. at end of three years for 95 subjects under age 15. (r . .963).

I have also computed the correlation between the initial I. Q. and the I. Q. at life age 15 (or with the last test when that was given prior to age 15). The coefficient is only a trifle lower than that shown in Table 1, namely .935. The correlation found by me for 428 repeated tests of normal children was almost exactly the same, namely, .933.

In further support of his contention that the I. Q. is of little value, the author cites the results of re-tests of borderline subjects by the N. Y. State Board of Charities.* From the original data of this report I have computed for the 49 subjects re-tested the agreement between first and second tests, which were separated in most cases by a year to a year and a half. The correlation, as shown in Table

*N. Y. State Board of Charities: Second Report on 52 Borderline Cases in the Rome State Custodian Asylum. 1915, pp. 32.

2, is .905. Only one marked case of disagreement stands out in the line of relation, and that is for a 5-year-old subject whose first mental age was 3.4 and who a year later tested at 5.3. Purely chance errors of this extent of course sometimes occur, particularly in testing young subjects, with whom it is often difficult to establish suitable rapport.

		Second I Q																
		35	40	45	50	55	60	65	70	75	80	85	90	95	100			
First I Q	105												1					
	100											1						
	95																	
	90																	
	85										2							
	80								1	1	1							
	75							2			1							
	70								1			1						
	65						1	3	3									
	60					3	3	1										
	55			1	2	2	4											
	50		1	2	2													
	45			3														
	40		1	1														
	35-39	1																

Table 2. Showing correlation between repeated tests. Rome (N. Y.) children, (r . .905).

Instead of the I. Q. having no value for prediction, as the author thinks he has demonstrated, the very data on which he bases this conclusion shows that even for feeble-minded subjects the I. Q. from three to five years hence can be predicted with a P. E. of less than 4 points, or that the final mental age which a subject will attain can, by use of the I. Q., be predicted from three to five years in advance with a P. E. of 4 to 6 months. The author (p. 53) explains his omission of statistical treatment on the grounds that "statistical devices, such as expressions of central tendency, coefficients of variability, and co-efficients of correlation obscure rather than clarify the results."

The author's main criticism of the I. Q. is based on the fact that the rate of intellectual growth, as indicated by two successive tests, is not predictable by its use. For example, if two children are tested and found to have widely differing I. Q.'s (one 70 and the other 90, say), re-tests a year or two later may disclose a larger mental age increase for the 70 I. Q. subject than for the other. Therefore,

the predictive value of the I. Q. is nil. However, the author overlooks the important fact that an I. Q., or any other kind of intelligence score, has a considerable probable error. The argument assumes that the score is a perfectly accurate measure of the thing it purports to measure. As Otis has shown,* the P. E. of a Stanford-Binet score for a group of adult delinquent and "hobo" subjects of average mental age 13 or 14 is approximately $5\frac{1}{2}$ months in terms of mental age, or about 3 points in terms of I. Q. With first grade school children of average mental age $6\frac{1}{2}$ years another of my students has found it to be about 3 months. Since these subjects are young, the P. E. in terms of I. Q. is again about 3 points. With another miscellaneous group the P. E. of I. Q. was a little less than 4 points. Accordingly, an I. Q. of 70 really means 70 ± 3 or 4.** Let us suppose that two ten-year-old children both of true I. Q. 100 were tested, and let us suppose the I. Q.'s found were 97 and 103, which would mean mental age scores of 9.7 and 10.3 respectively. Let us suppose that a year later their true I. Q.'s are still 100, and their true mental ages are 11, but that re-test brings a reversal of the error, giving 103 and 97 I. Q. respectively. Their mental age scores would now be 11.33 and 10.67. The first subject would appear to have gained 1.67 years and the latter .33 of a year. Doll's argument would assume that one had gained five times as rapidly as the other. The point is, of course, that no intelligence test yet devised can legitimately be used for measuring the rate of growth over relatively short periods, since the P. E. of a mental age score is itself 25 to 50 per cent varying with age of the normal amount of growth for a year. If the author were to test a group of subjects on two successive days, as one of my students has done, he would find almost as large and as frequent I. Q. changes as his data show for tests separated by a year. If in such an experiment a subject were found to have gained a half year in mental age, surely Dr. Doll would attribute this to an over-night spurt in mental growth.

It is the same fallacy which accounts for the author's argument that the I. Q. is misleading because, being a function of the entire life age, it "irons out" significant mental growth changes by making

**J. of Educational Research*, March, 1921. The publication of this study, which was made in 1916, was delayed by the war.

**The author's mental ages, being averages of two or more tests, would have a somewhat smaller P. E. than this, probably a little over 2 points.

them a fraction of a large unit.. On this ground the author argues that an I. Q. change of even 5 points is very significant, indicating as much as a halving or doubling of the mental growth rate in the interval between the tests. Thus far the author's fallacious reasoning may be accounted for by his neglect to take the probable error into account. Even apart from the probable error, however, the argument is unsound. What it amounts to is this: the I. Q. is blamed because even a 50 per cent increase or 50 per cent decrease in growth *rate* does not in a short interval greatly alter its value. Why should it? If this could occur the I. Q. would no longer be an index of brightness at all. A 50 per cent gain or loss in *rate* of mental growth for one year, even if such gain or loss were genuine and not a mere accident of score unreliability, would not greatly alter a 10-year-old's brightness status with reference to the norm for his age. It is of course just this brightness status for which the I. Q. is intended to serve as an index.

The author takes exception to my statement that the fairly constant variability in the I. Q. distribution at different ages contradicts the traditional view that variability in mental traits increases toward adolescence. His criticism is based on the fact that there is greater age overlapping in intelligence as maturity is approached and that this is only an expression of the increasing variability.

However, in common with Bobertag, Stern, Kuhlmann and others, I have myself pointed out this increase in age overlapping. My statement regarding variability was of course based on the fact that I consider mental age a misleading unit in which to express variability? Surely, the child of 12 years with a mental age of 11 does not vary as much from normal as the 3-year-old who has a mental age of 2.

The author also criticises my statement that "the mental age of a subject is meaningless if considered apart from chronological age"* in such a way as would lead the reader to assume that I consider mental age of little significance as compared with the I. Q. He proceeds, as though refuting my view, to show that it is mental age, rather than I. Q., which determines the school grade or the kind of vocational employment which is suitable to a given subject at a given time. In view of the fact that I have written two books largely to show the value of mental age as a basis for school grading, and have

*The Measurement of Intelligence, 1916, p. 68.

many times pointed out that the I. Q. is useful as an index of brightness, but not (apart from age) of ability level, the author's criticism hardly seems fair.

Again (p. 16), "Terman maintains that mental growth develops at a constant rate for all degrees of brightness and dullness (except idiots and low-grade feeble-minded)". Here the author refers to my book, *The Intelligence of School Children*. If he will read this again (Chapter 9) he will see that I simply state what is true for the data I offered, namely, re-tests of 315 children of whom only 31 were below 80 I. Q. On p. 147 of the same book I expressly state that feeble-minded children testing below 60 may be less likely to hold their own than those of milder degrees of defect. On p. 150 I warn the reader against accepting the I. Q. as infallible and state that "in pathological subjects it may undergo large fluctuations." On p. 154 I state that we could hardly expect the I. Q. to remain absolutely constant even if it were based upon a perfectly accurate scale, which I expressly pointed out we do not have.

Other inaccuracies include the following:

P. 6, "It appears from Terman's tables of standardization statistics that he did locate the single tests according to the general principles employed by Binet" (i. e. by the 75 per cent rule). Here the author refers by reference number to my monograph on the *Stanford Revision*.^{*} However, in Table 43 of this monograph the per cents holdings for the *Stanford Revision* are explicitly shown to decrease gradually from an average of 77 per cent for the tests of year 4 to less than half this amount at the upper end of the scale. Elsewhere (p. 10) the author, contradicting his other statement, gives me credit for ignoring the 75 per cent rule, but asserts that the results are nevertheless the same. Still later (p. 124) he presents a table which shows the results are far from the same, a Goddard mental age at 5 or 6 years being shown to be nearly a year higher than my own and at 10 no higher.

P. 7. "It [the Binet scale] is based on the arbitrary assumption that increments in mental age from year to year are equal in amount." If "equal" here means anything it means equal in terms of some kind of absolute units. Of course we have no such units and are not likely to have soon. As a matter of fact the Binet type of scale does not necessarily presuppose equality of mental age steps.

^{*}Warwick and York, 1916.

P. 70. [There is] "a tendency for the lower mental ages or the more retarded subjects to show larger rates of increase than the higher mental ages or the less retarded subjects." I have calculated, as well as I could from the author's data, the relation between mental age increase and I. Q. for the 67 subjects who were followed for as much as three years before life age $15\frac{1}{2}$ (or thereabouts). The results are as follows:

I. Q. at initial test	0-29	30-54	55-88
Number of cases	7	26	33
Approximate average I. Q.	20	41	.70
Expected gain in 3 years.....	.6 year	1.2 year	2.1 years
Average gain found2 year	.8 year	.9 year
Ratio, found to expected.....	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{3}{7}$

The important fact here is that those below 30 I. Q. tend to develop at $\frac{1}{3}$ the expected rate, those from 30 to 54 I. Q. at $\frac{2}{3}$ the expected rate, and those above 55 at $\frac{3}{7}$ of the expected rate. The last fraction however, is too low, due to the fact that the mental age range from 7 to 10 on the Goddard Revision represents more nearly four years of mental growth than three.

P. 101. "The individual average annual rate of development for these subjects in most cases is less than 20 per cent. This in spite of the fact that most of the first I. Q.'s of these subjects [those under 15 years at initial test] range above 50." It appears, however, that nearly half of the first I. Q.'s. were below 50, and that the average annual rate of development for those above 30 was for the middle group two-thirds of the expected and for the high group three-sevenths (perhaps actually four-sevenths) of the expected.

Dr. Doll also gives the results of Miss Gillingham's retests of 35 superior children with I. Q.'s above 110. With regard to growth irregularity and I. Q. constancy the author's conclusions from these tests are in line with those based upon tests of the feeble-minded and as little supported. However, he presents this hypothesis that the I. Q.'s. of superior children tend to increase rather than decrease. "We have found [for the 35 superior children of the ages 10, 11, and 12] the average rate of growth to be distinctly higher than the average age I. Q., showing a tendency for the superiority to increase." I find from his data that the central tendency of I. Q. change for these subjects was —2 points. I also find for these subjects a negative correlation of —.474 between I. Q. at first test and amount of im-

provement. That is, the brighter the superior child the less likely is the I. Q. to increase, which is contrary to the author's hypothesis. Of the 27 below 135 I. Q. at first test, exactly two-thirds showed an increase; of the 8 above 134, only one-fourth. With a highly selected and relatively narrow range group such as we have here, a negative correlative between first and later tests would be a natural effect of the probable error of the scale. That is, the higher the initial I. Q., the greater the chances that it is in error on the positive side; the lower the initial I. Q., the greater the chances that it is lower than it ought to be. However, the author's data on superior children are too scanty to warrant any conclusions whatever, and in so far as they indicate anything it is the reverse of what his hypothesis lays down.

In general, if one would know what the author's data really show it is always necessary to determine this for one's self from his tables of results. His own conclusions are so often either contrary to his facts or else irrelevant to them that verification is always necessary. One is tempted to offer the injunction *caveat lector*.

(To be continued in October.)

DEPARTMENT FOR DISCUSSION OF RESEARCH PROBLEMS



Conducted by LAURA ZIRBES



This department has a two-fold function. It aims to serve research workers as well as educators, whose work brings them in close contact with children in the schools. It hopes to accomplish this service by suggesting research studies, which will meet well-defined school needs.

In order that this service may be real and effective, the co-operation of research workers and school people is desired. Correspondence with reference to the following questions will be considered in selecting topics for future discussions.

- a. Which of the studies proposed would help you to solve a practical problem?
- b. What topics might well be added to this list? Replies may be addressed to: Miss Laura Zirbes, 646 Park Ave., New York City.

What is the maximum age and ability range of an effective pupil group? What range is most desirable in a class? Within what limits is difference in age or ability immaterial? Is the range the same for all school subjects, or different for content and tool subjects, for problem solution, construction projects and drill? What ages or grades combine best in working groups? How great is the difference in efficiency among groups of different ranges?

These are important questions for rural schools. There are some two hundred thousand one teacher schools in the United States. The total enrollment of each tends to be small, though all the elementary grades may be represented. For economy of time, and for social ends as well, it is desirable to make fewer groups than the number of grades calls for. To what extent is such grouping desirable also from the standpoint of efficient instruction?

FANNIE W. DUNN, Teachers College.

The effect of textbooks on the outcomes of instruction. In another part of this issue reference is made to a study of the effect of textbooks on methods of instruction. This would indeed be a fruitful topic for further investigation and research, if the making of school texts is to proceed scientifically and if the hope for improved instruction is to be based on improved texts.

The rigid adherence to school texts demanded of high school pupils in mathematics, history and literature, and in the sciences is

certainly not warranted by the psychological validity of the texts in question. A similarly rigid adherence to book-made plans in the elementary school is incompatible with the needs of particular classes and individuals and interferes with the spontaneity and flexibility which characterize teaching based on the actual experiences, interests and purposes of the pupils in question. The teacher who notes the reactions of her pupils, and tactfully endeavors to adjust her work to the situation is not as rare as the one who combines with this type of leadership a clear vision of the aim and purpose of each lesson in relation to general educational outcomes, and a knowledge of the psychological conditions necessary to the realization of these ends. Research along these lines will be of more practical value than assistance in the actual outlining of lesson plans for class-room use.

At present, textbooks differ in organization and content, form and purpose, depending on the point of view and experience of their authors, and the demands of the publishers and the public. Whole sets of readers are recommended and adopted because of their extensive study helps. Investigation shows that the study helps are not used. Books and lessons made to teach the mechanics of reading are used as texts in language or literature. Music is given place in the curriculum for its aesthetic value but is taught by the aid of books which stress formal and technical drills and even make songs subservient to these purposes. Subjects retained in the curriculum for supposed disciplinary values are taught by strict adherence to texts or methods which have never been subjected to experimental evaluation and whose outcomes are of doubtful educational value.

Thus textbooks often prevent purposes, hamper learning and perpetuate obsolete educational ideals; or, by partial adjustments of content and method, seek to become sufficiently expurgated and imbued with modern ideas to pass muster.

No doubt, the publication of texts is more remunerative than the investigation and evaluation of bodies of material and methods of presentation as means for the attainment of worthwhile purposes. There is, nevertheless, great practical value in the impartial analyses of current methods and results. Permanent contributions to educational practice and literature can hardly be compiled without careful studies covering the social worth of their content, the psychological

implications of their organization, and the degree to which they serve adequate educational purposes. Only a small number of texts now in use are the result of scientific investigation and research. No doubt, these will maintain their standing and effectiveness, and raise the standard of other materials offered for publication.

L. Z.

NOTES ON ARTICLES IN EDUCATIONAL PSYCHOLOGY IN CURRENT ISSUES OF OTHER MAGAZINES

Reported by CECIL COLLOTON,

Department of Educational Psychology, Lincoln School of Teachers' College.

EDUCATIONAL TESTS.

The Measurement of Language: What is Measured and its Significance. Ernest J. Ashbaugh, Journal of Educational Research, 1921, June, 32-39. Analysis of various scales for the measurement of language; their merits and their limitations.

Scale of Attainment No. 2—An Examination for Measurement in History, Arithmetic, and English in the Eighth Grade. S. L. Pressey, Journal of Educational Research, 1921, May, 359-369. Description of the examination; statement of norms for the examination as a whole and for each test; and discussion of the usefulness of the scale in comparing graduation standards.

A Handwriting Scale for the Pupil. Frank Freeman, Elementary School Journal, 1921, June, 755-761. Description of a series of scales to enable the pupil to measure his own handwriting. Details of construction of scale.

Graphical Representation of Grades of High-School Pupils. Elbert Allen School Review, 1921, June, 467-471. Description of a set of three cards for graphic representation of test grades, (1) card for class graph; (2) individual score card; (3) card for comparison of 4 specific types of graph.

INTELLIGENCE TESTS.

Fuctuation of Intelligence Quotient. S. C. Garrison, School and Society, 1921, June, 647-649. Results of retests on 62 children in the Peabody Demonstration School.

The National Intelligence Tests. Guy M. Whipple. Journal of Educational Research, 1921, June, 16-31. Brief description of the history of the tests, the aims of the makers, the criteria that were observed, and the results that are being obtained.

The High Cost of Testing. S. L. Pressey, Elementary School Journal, 1921, June, 771-777. Discussion of three practical criteria to be considered by superintendents in selecting tests.

The Intelligence Test and the Teacher. Otto W. Haisley, Elementary School Journal, 1921, May, 703-707. Results of an Intelligence Test given to all teachers in the school system at Niles, Michigan.

Intelligence Tests in Classifying Children in the Elementary School. Charles Fordyce, Journal Educational Research, 1921, June, 41-43. Study of the results of the Haggerty Intelligence Examination in comparison with the school grades and estimates of teachers in the case of 1078 pupils in the elementary grades at Lincoln, Nebraska.

The Reliability of Test Scores. Truman L. Kelley, *Journal of Educational Research*, 1921, May, 376-379. Critique of nine methods used in measuring reliability of tests with a recommendation for establishing a standardized procedure.

Suggestions Looking toward a closer contact with Practical Problems in Work with Educational Tests. S. L. Pressey, *School and Society*, 1921, June, 710-716. A caution against the indiscriminate use of statistical procedure without reference to its application to the data and problems in hand.

The Intelligence Examination for High School Freshmen. Ira J. Bright, *Journal of Educational Research*, 1921, June, 44-55. The Terman Group Test of Mental Ability in comparison with teacher's marks in Latin, English, Algebra, and Handicraft Subjects. Using the test as a basis for organization of class groups.

Standardizing Tests for Vocational Guidance. James Burt Miner, 1921, June, 629-633. *School and Society*. Making tests useful for vocational placement by (1) the measurement of occupational types and (2) the measurement of the most stable workers within an occupational group.

Reclassification of Children on Basis of Tests in Port Chinton Schools. A. F. Meyers, *Journal of Educational Method*, 1921, Sept. 24-25. Classification of children on basis of intelligence and achievement tests, and establishment of mid-year promotions.

The Freeman-Rugg General Intelligence Tests as an Aid to Economy in School Administration. Ray H. Bracewell, *The School Review*, 1921, June 460-466. The classification of Freshmen pupils in ability groups on basis of Freeman-Rugg tests in the Burlington, Iowa High School.

Psychological Clinics in the United States. Leta S. Hollingworth, T. C. Record, 1921, May, 221-225. History of the Psychological Clinic and its present status in the U. S.

Some Results from a Testing Program in Idaho. I. N. Madsen, *School and Society*, 1911, June, 668-671. Results for Idaho Schools in Haggerty Intelligence Examination, Monroe's Silent Reading Tests, and Monroe's Reasoning Tests in Arithmetic.

Group Mental Testing in Altoona, Pa. Caroline E. Meyers, Garry C. Meyers, S. H. Layton, *School and Society*, 1921, May, 624-628. Results of testing 6,774 children of the elementary schools of Altoona, Pa., with the Myers Mental Measure.

A Program for Lowering the Percentage of Failures. Harlan C. Hines, *School and Society*, 1921, May, 582-584. The use of Terman's Group Test of Mental Ability in Los Angeles public schools.

Norms for the Sequin Form-board. Based on the averages for three trials. J. E. Wallace Wallin. *Journal of Delinquency*, 1921, May, 381-386.

Presenting Educational Measurements so as to Influence the Public Favorably. Carter Alexander, *Journal of Educational Research*, 1921, May, 345-358. Talking points on measurement for convincing taxpayer of educational needs and securing better school support.

Minor Studies from the Psychological Laboratory of Indiana University. VIII. A Preliminary Investigation of General Prognosis; i. e., General Intelligence." *Journal Applied Psychology*, 1921, Mar., 78-84. Comparison of scores made by Junior High School class on a group intelligence tests and the marks made by the same children over a year later to determine prognostic efficiency of the tests. IX. Further data with regard to sex differences. Sex differences in mental and emotional traits as determined by various tests.

Educational Guidance and Tests in College. Stephen S. Colvin, *Journal Psychology*, 1921, March, 46-56. Description of a series of tests for use in a system of educational advice and direction for its students based on psychological tests.

A Test Series for Journalistic Aptitude. Max Freyd, Journal of Applied Psychology, 1921, March, 46,56. Description of a series of tests for use in vocational selection and guidance in the field of journalism.

Tests in Industry. Morris S. Viteles, Journal of Applied Psychology, 1921, Mar., 57-63. Need for tests for selection of workers in industry and difficulties attending the development of such tests.

The Problem of the Unselected Group in the Standardization of Tests. S. L. Pressey, Journal of Applied Psychology, 1921, Mar., 64-71. The problem of obtaining unselected, representative groups of cases for the determination of norms on the various types of test.

LEARNING IN THE SCHOOL SUBJECTS.

Analysis of Learning Processes and Specific Teaching. Charles H. Judd, Elementary School Journal, 1921, May, 655-664. Following up tests by analysis of results, study of particular cases, and specific teaching based on intensive analysis.

SPECIAL REVIEW OF MRS. BURGESS' MONOGRAPH ON SILENT READING.

BURGESS, MAY AYERS. *The Measurement of Silent Reading*. New York: Russell Sage Foundation. 1921. Pp. 163.

One of the major virtues of Mrs. Burgess' work is that it uses the analytical method and defines, with some degree of detail and exactness, what is being measured. It has become the custom in recent years for authors of tests to proclaim with complete complacency that they do not know what they are measuring. The criterion of validity which has been accepted by such makers of tests is a high correlation with some other test equally vague in its purposes and its object of attention. The result of this cumulative indefiniteness is that it is easy to get on the open market a great many hastily devised tests that have been handled and rehandled by formal statistical methods, but are altogether nondescript with regard to their value as instruments of educational diagnosis.

The vagueness that has characterized tests has appeared also in the definition of school subjects. What has not been included in the last few years under the term "reading"? All sorts and kinds of exercises which use printed words have been called tests in reading, whether they measure the mechanics of reading or the most abstract forms of reasoning. On the other hand, the importance of reading has often been overlooked in such tests as the so-called reasoning test in arithmetic. Authors of reasoning tests seem to assume that every child can get from the printed page the ideas involved in solving an arithmetical problem, even when the statement of the problem is intricate enough to baffle the reading power of an adult.

Mrs. Burgess very properly challenges the work of the vague testers and sets an example which ought to give pause to the reckless publication of half considered devices for measuring mental processes. She points out that every test should aim at some particular point and not try at one stroke to do everything. She then tries out one after another of the devices which seem to her to fit her particular purposes and discards the tests which are not satisfactory, until she develops a perfected instrument. All this is done without hiding behind a smokescreen of higher mathematics that is intended to frighten off the critic and provide a shelter for vagueness and lack of insight.

Such critical comments as can be added to the foregoing approval of what has thus been accomplished in Mrs. Burgess' book ought perhaps to be postponed until the full impact of her discussion of tests has had time to be felt. At the risk of making a mistake and with the hope of promoting rather than in any way hindering the progress of analysis, I shall venture to point out some of the detailed difficulties which I find in the book.

The list of factors controlling silent reading, given on pages 37 and 38, evidently includes a mixture of many different kinds of items. Some are psychological, such as attention span; others are wholly objective, as uniforming of print. Some are very easy to keep constant; these are the object factors. Others are much less accessible to the experimenter, and it is by no means as certain that the test has succeeded in keeping them constant. Indeed, it is the belief of the present writer that attention span is always a variable. Whenever we work in the psychological laboratory we find that we must take into account fluctuations of attention. It is never possible to get two individuals with the same span of attention, however constant we make the external conditions. It would seem wise, therefore, in cataloguing the constants in a reading experiment to classify external and internal factors separately, and to interpret the final result with due regard to the impossibility of holding subjective factors constant in the same sense in which we can control objective factors.

The foregoing discussion will make clear the reason why there is objection to the statement made by Mrs. Burgess on page 61. She writes, "The process by which the essential characteristic of constancy is obtained in educational measurements is the one used in physical measurements. It consists of distinguishing the possible controlling, varying factors; devising means for holding them all constant save one; and measuring that one. This is the law of the single variable."

The objection to this formula is that in matters psychological it is the organized whole which is important rather than any single factor. The history of reaction-time experiments is instructive in this case. It was assumed at one time that a simple reaction is a detachable part of a complex reaction. The subject of the experiment was asked to lift his hand at a given signal and the time of this simple reaction was measured. Later the same person was

asked to react under more complicated conditions. He was to lift his right hand if the signal was of one type, and his left if it was of another. The time required for this choice reaction, as it was called, was longer than the time required for the simple reaction. It was argued that the choice reaction was made up of the simple reaction, plus the factor of choice. The argument ran on this wise; in both complex and simple reactions the same eyes or ears receive the signals, the same hand responds, the difference is the one item of choice.

Experimenters worked long and arduously on the assumption that their arguments were valid, but their results were full of curious inconsistencies and finally led them to see that a complex reaction-time is not a simple reaction plus an additional factor. A complex reaction is a new organized whole. It cannot be torn into parts as can a physical compound. It is what it is by virtue of its organization.

This lesson from the history of psychology should be taken seriously as a guide to all who undertake psychological analyses. The principles of such analysis cannot be borrowed from physical science, and the factors sought are of a character different from the physical factors involved in an experiment in the natural science laboratory.

The elaboration of this criticism is not intended to suggest scepticism as to the validity of Mrs. Burgess' test. Her analysis has gone far enough to make her instrument of diagnosis very useful in a practical way and to render it much more definite in purpose than most tests. The refinement of her method and the final solution of the reading problem wait, however, for the further productive analysis of the reading situation. The plea which is made here is for more analysis, guided by experience collected through such studies as Mrs. Burgess has made.

CHARLES H. JUDD.

In this important monograph on silent reading, Mrs. May Agnes Burgess has rendered at least three services: (1) She has made a useful exposition of the laws of a scientific procedure in the construction of reading tests; (2) she has contributed a suggestive analysis of reading as a function, and (3) she has set up an admir-

able sample of the experimental and statistical study of an instrument that should precede, rather than follow, its publication for general use.

In the discussion of scientific methods as applied to the construction of reading tests, Mrs. Burgess advances, as the assumption of first importance the "Law of the Single Variable." "It consists of distinguishing the possible controlling, varying factors; devising means for holding them all constant save one; and measuring that one. This is the law of the single variable." P. 61. The spirit of this dictum is admirable, but the wording is too inflexible. Students of the history of scientific methods are familiar with this law as it is applied in the physical and mental sciences and will agree with Mrs. Burgess that absolute control of all variables save one is the *ideal* of scientific procedure. In the biological sciences it is not always possible to follow it rigidly. It is really not necessary and sometimes not possible, however desirable, to hold all variables constant. It is often only necessary or feasible to *take all variables into account*. In the physical sciences, as Mrs. Burgess points out, it is usually best and usually possible to control all variables save one, but in dealing with integrated human reactions, it is frequently impossible. Our method is still scientific, however, if we can observe and measure the variables. In Gray's oral reading test, for example, speed cannot be held constant when accuracy is being measured, but if speed is measured and allowed for, the demand of science is met. It is sometimes possible and proper to sidestep a variable, as, for example in the Thorndike test, by giving a maximum of time when the purpose is to measure power of comprehension freed of the mechanics of reading. Doubtless it was its Mrs. Burgess' intention chiefly to indicate the necessity of thus taking into account the several variables and she has, in fact, given an admirable sample of an effort to analyse silent reading into elemental factors. The necessity of controlling the vocabulary, phraseology, sentence structure, the motor reactions of carrying out the directions, the length of the paragraphs, the interest to the child, etc., and of eliminating arithmetical information, memory and other factors not necessarily involved in reading ability, is effectively considered. It is no reflection on the theoretical discussion if all of the variables involved in the author's own test are not perfectly controlled. The various paragraphs were standardized

for difficulty by selecting those whose directions were fulfilled by the same percentage of children. By the use of this single criterion it was possible for the time required to read and the time required to draw the supplements to vary from paragraph to paragraph. As a result of an experiment this was found to be the case.¹ It should be added, however, that while these factors were not experimentally controlled an effort was made to equalize them empirically. If a test constructed with such care as the Burgess shows such defects, what of most others?

The monograph calls our attention anew to a really neglected line of research; that of the interrelations of the several variables or "dimensions" of functions such as quality, accuracy, difficulty and speed. What is the effect upon quality of handwriting when speed is varied? How can we control the one while measuring the other and if we cannot, what allowances are to be made per unit change in either. In reading, writing, composition, etc., the variables are related in still greater complication.

A most commendable feature of the work with the Burgess Scale is the fact that statistical studies of its validity and reliability were conducted before the test was put on the market. We have more than the author's opinion that the test does measure reading and a notion of how well it does it. In connection with the question of "reliability" or "consistency," Mrs. Burgess has presented an unusually clear account of the limitations of our conventional statistical methods and interpretations. The co-efficient of reliability, dependent upon the correlation of repeated tests with the same instrument "may constitute in some measure a basis for valid criticism of the test, but in the main (it appears) to reflect a real and inevitable variability of human performance. The important fact to remember about such scores is that they may vary from day to day and still be actual true measures of ability on each occasion. Under such conditions the fact that the scores vary from trial to trial does not reflect any inaccuracy or inadequacy of the test," p. 131. It does, however, indicate an inadequacy of the test results for practical purposes and clearly indicates that further study of the test is necessary. It is true that the co-efficient does not itself indicate the "cause" of variability. We may blame the

*The data are to be presented in the October issue of this journal.

child, if we please, but we cannot change the child whereas we can change the test. It may be that the test should be lengthened, or that the units should be more carefully equalized or made more fine, or—well, only further experimental study will disclose what changes may be profitably made. At any rate, it is certain that a test to meet present-day requirements must yield highly consistent results.

Perhaps the most fortunate outcome of Mrs. Burgess' vigorous discussion will be its effect upon our attitude toward new tests. With this admirable example of scientific analysis and research before us, it will be quite inexcusable for anyone to throw on the market a new test whose validity and reliability has not been very thoroughly determined.

ARTHUR I. GATES.

Whatever else may be said of the latest educational monograph issued by the Russell Sage Foundation, there is no gainsaying the fact that "The Measurement of Silent Reading" and the accompanying Scales merit the careful consideration of serious thinkers and workers in the field of educational measurement. Dr. Burgess set for herself a task of no mean proportions and her book is an exceptionally clear-cut and well-organized record of her procedure, and an interesting discussion of some of the principles and laws underlying educational measurement. The *raison d'être* of every step in the construction of the scale is given and the alternatives or methods used by other workers are critically examined and evaluated. The limitations of existing scales are set down at length. Several chapters are given to the discussion of the law of the single variable and its implications in the field of educational measurement. Dr. Burgess lists variables of quality, of difficulty, and of amount, and maintains that the student of educational measurement must consider first which of these three he will attempt to measure, and then use the type of scale adapted to the measurement of that variable, excluding from the test or rigidly controlling other variables, so that comparisons based on measurement of one variable may not be adulterated by intruding factors.

A scale like the Ayres Burgess Silent Reading Scale certainly facilitates comparison inasmuch as the units of the score are equal quantities of a single variable, measured under conditions in which

other factors are carefully controlled. This makes it an effective survey instrument or device for sampling the abilities of large groups.

While it is thoroughly scientific to construct a test with due regard for the law of the single variable, the practical and scientific value of a test depends even more upon its reliability and validity than upon its simplicity and statistical qualifications. The question of validity is not discussed in the monograph. The test material is sufficiently unlike that used in most other school exercises to make this consideration significant. Furthermore, the very fact that the whole test is of approximately the same difficulty and requires but one type of reaction, leads one to wonder how much dependence may be placed on the scores in gauging the reading abilities of individuals. There are so many specialized reading abilities necessary to the proper performance of school duties and the satisfaction of the responsibilities and privileges of society. Diagnostic analysis of reading deficiencies points the way to specific training to the end that pupils acquire habits and sets which function in response to the varying demands of particular situations. General advice based on performance under one set of conditions is of doubtful value. The solution is a group of tests covering the specific reactions demanded in a representative group of reading situations, each one meeting the other scientific requirements so ably set forth by Dr. Burgess in her monograph.

Laura Zirbes.

NEW PUBLICATIONS IN EDUCATIONAL PSYCHOLOGY AND RELATED FIELDS OF EDUCATION

1. *Three psychological studies of school children.* The literature dealing with the superior child is growing constantly every year, and two of the studies to be reviewed here deal with that subject. These two studies, however, show a great contrast in method, inasmuch as the one seems to revert to subjective methods for selecting superior children, while the other bases the selection upon objective mental tests.

The short monograph by Badenes¹ is, therefore, very disappointing to the psychologist. It lists nine pages of items or traits, mental, emotional, physical and so forth, supplies a chart on which the teacher is supposed to evaluate or describe these traits and that seems to be about all. The great value of mental tests seems to be disregarded, and the recent progress in the construction of rating scales, very pertinent to the author's method, has been entirely ignored. The historical references in the introduction center around Stern and Meumann, while the valuable contributions of the United States in the work with superior children are scarcely mentioned.

Superior children only enter incidentally as one of the groups of exceptional children studied by Gesell.² The book reports chiefly the results of a rapid survey of 24,000 elementary school children in New Haven. The exceptional children were reported by the teachers and physical as well as mental defects were listed. There were 370, or about 1.5 per cent reported mentally deficient. In contrast with this only 45 were reported mentally superior, and the author comments on this fact. It strengthens the reviewer's conviction that mental tests are more needed for the selection of the superior than they are for the inferior. The great discrepancy between the numbers of superior and inferior children reported is, in the present instance, to some extent due to the fact that the New Haven survey emphasized the question of mental deficiency. All the children reported to the teachers as mentally deficient were

¹Badenes, J. E. *The First Practical Steps in Selecting Gifted Children in a Large City School.* New York. 1921. Pp. 22.

²Gesell, A. *Exceptional Children and Public School Policy.* Yale University Press. 1921. Pp. 66.

given Doll's short Binet and some other tests. Some of the chief results are given on figures 12 and 13, although there seem to be serious discrepancies as shown on the figures. Furthermore, the comparison of the distribution curves on figure 12 is vitiated by the difference in the totals of the two groups compared. The results of this survey are used to make specific and valuable recommendations for the local situation and a model program for the community care of mentally deficient school children is presented.

The last of our monographs by Town³ is an intensive study of 52 children during the first months of their school life. There are very few children in the world who have been so thoroughly measured, physically, anthropometrically, and mentally, as have the 52 children of this study. As many as 51 mental tests are shown on the psychological profiles, and because of the great number of tests these profiles become exceedingly difficult to read. Profiles based upon medium percentiles of groups of allied tests would have been more illuminating. Apart from the Binet and the Stanford Revision, only two tests, the Knox-Pintner Cubes and the Porteus Maze, were interpreted in the light of previous standardizations. In both cases these previous standardizations do not seem to suit the author and she remarks ambiguously that neither of the tests show much correlation with chronological age. This is a very peculiar statement to make after testing only 31 five-year olds, 12 six-year olds and one seven-year old, all of whom were children in the first year of school. An inverse correlation with chronological age is to be expected with such a group, and indeed on the cube test we find that the median at age five is two years advanced, at age six exactly age six, and at age seven (one case) two years retarded.

R. PINTNER.

2. *Educational Psychology by a French Writer.*—The eighth edition of Claparède's work was issued in 1920, the first in 1905. An English translation of the fourth edition appeared in 1911. The present volume¹ is very much enlarged since the fifth edition of 1915; but, with very few exceptions, the bibliographical references

³Town, C. A. *Analytic Study of a Group of Five and Six-Year-Old Children*. Studies in Child Welfare. Univ. of Iowa. Vol. I, No. 4. May, 1921, Pp. 87.

¹Claparède, Ed. *Psychologie de l'enfant et Pédagogie Expérimentale*. Geneva: Kundig, 1920. Pp. XL-571.

have not been brought up to date. The principal changes are the omission of the chapter on fatigue and the very great extension of two other chapters, so that the work in its present form is more than twice the size it was ten years ago.

Let no one be misled by the title into expecting a treatise on child nature. It is rather an argument for the value of experimental education, telling why we should study children, how to do so, and who has done so, with a minimum of results found except as they illustrate points in method. At the beginning is given an excellent outline of the history of child study, with a summary of the work done in different countries to date. Then follows an exhaustive analysis of the problems with which the science has to deal, and the methods by which it attacks them. The last chapter, on mental development, takes about one-quarter of the book only, and fully one-third of that is devoted to the topic of play. After discussing the various theories of play the author announces his own view of its function: that of allowing to the individual a realization of the self by following a line of greatest interest when more serious activities do not afford scope for that for the time being. Classifications of play follow, and a brief description of the dominant interests at different ages, with a plea for their greater utilization in school tasks. Evidently in sympathy with Dewey's writings he would surely advocate the introduction of the project method to the schools of his country.

M. T. WHITLEY.

3. *A Helpful Summary and Interpretation of Experimental Evidence on Silent Reading.*—O'Brien's recent book, *Silent Reading*,² is undoubtedly the most helpful single summary and application of contemporary psychology in the field of school reading that has been published. It does two things very well indeed. First, it discusses, in language that the school administrator and teacher can understand, the previous scientific investigations which have dealt with movements of the eyes during the reading process; with comparisons of oral and silent reading; with the factors which contribute to rapid silent reading. (He finds the most important to be, (1) practice in rapid silent reading; (2) decrease of vocalization; (3) training in perception; (4) character of subject-matter; (5) habits of eye movement; (6) purpose for which subject-matter is read; (7)

²O'Brien, J. A. *Silent Reading*. New York: MacMillan, 1921. Pp.

concentration of attention; (8) ability to grasp the meaning of contents, etc.) with the difficult and necessary reduction of vocalization; and with training in increasing perception. I know no better statement of the present investigational status of these matters than Dr. O'Brien gives.

On the side of positive and constructive training of children O'Brien reports his very extensive experimental investigation of methods of obtaining rapid and effective silent reading in 40 school classes. (Grades III-VIII.) As a result of this investigation and by interpreting contemporary thinking as it concerns the teaching of reading, he presents definite recommendations for the development of rapidity in rate of reading, for reducing vocalization and for increasing perceptual ability.

H. O. R.

4. *A Book on Mental Tests by An English Writer*.—"Why is it that America has been moving so rapidly in the matter of mental tests while England has almost stood still? The answer is simple: Speaking generally, Americans believe in psychology, but Englishmen do not. When America entered the war, one of the first things she did was to mobilize her psychologists. The war was nearly over before England discovered that psychologists were of any use."

This quotation is from a recent English book,¹ in which the author undertakes a task which would be unnecessary today in America, namely, to persuade the teacher to "believe in psychology."

In the first chapter he replies to those who instinctively dislike the idea of bringing measurement into education, and gives a brief historical account of the development of Mental Tests. This is followed by a discussion of the subject of general intelligence, and of the work of Binet. Though the Terman revision is accessible in England and is commended by Mr. Ballard, he quotes in his text the translation of Binet's tests which was made by Mr. Cyril Burt in consultation with Binet's collaborator, Dr. Simon. Mr. Burt rearranged the tests in order of difficulty, and made the age-assignments in accordance with the results of his experiments with a large number of children in the London Elementary Schools. Unfortunately, the exact number of children is not stated.

It is interesting to compare the different ages to which Burt and Terman assign the same tests. In several of the lowest tests, the

¹Ballard, P. B. *Mental Tests*. London: Hodder & Stoughton. 1920. Pp. IX+235.

American child is in advance of the English, while later he seems to be out-stripped; e. g., in the repetition of numbers the age standards are as follows:

Repetition of 3 numbers	Age 3	Age 4
Repetition of 4 numbers	Age 4	Age 5
Repetition of 5 numbers	Age 7	Age 6
Repetition of 6 numbers	Age 10	Age 9
Repetition of 7 numbers	Age 14	Age 11

In giving this test Burt read the numbers at the rate recommended by Binet, i. e., two per second, while Terman's directions give "a slightly faster rate than one per second."

The American is evidently quicker in verbal repetition all through his childhood, as at the age of three he can repeat six syllables, an accomplishment of the English four-year old; at five he can manage 13-15, while the English child can repeat only 10; at six, he repeats 16, the English seven-year standard.

It is at the higher age levels that the English boy or girl outstrips the American, e. g., the fifteen-year old does one of Terman's "Average Adult" tests and two at the "Superior Adult" level. The only test for English fifteen-year olds that Americans can do earlier is the question, "What are the three chief differences between a King and a President?"—a test obviously more suited to Americans.

Both Mr. Ballard and Mr. Burt feel the inadequacy of Binet's tests, and also of Terman's additions, for the discovery of the brightest children. For this purpose Mr. Burt has drawn up a series of 50 Reasoning Tests for ages 7 to 14, (re-printed in this book). They are individual oral tests, and the score can easily be changed to a Mental Age, and a Reasoning Quotient obtained. They are ingenious little puzzles, involving the application of thought to the ordinary affairs of life, and almost as applicable to America as to England,—in fact, they have recently been tried with success in this country.

The second part of the book is concerned with Educational Tests, prefaced by a clear and useful account of statistical distribution and dispersion. The tests and age-standards given here, in reading, spelling and arithmetic, show that England is beginning to strike out for herself along the path opened up by Thorndike, Courtis

and others. Ballard acknowledges his debt to the American school, and appeals again and again for *age* standards instead of *grade* standards, the latter being unintelligible to English readers.

The book will undoubtedly serve its purpose,—to popularize the subject with English teachers, and to disprove the allegation of the British press that mental tests are nothing but “new American fads.”

E. I. NEWCOMB,

Institute of Educational Research Teachers' College.

5. *Physical Growth of Children.*—This monograph¹ offers a complete survey of investigations on the physical status and growth of children, together with a report of the technique and results of the extensive work of the Iowa Child Welfare Research Station under the direction of Bird T. Baldwin.

Part I, dealing with the anthropometric instruments and methods, will serve as an excellent manual for class work. Part II, gives the data on the weight and height of infants, including table of norms, correlations and 400 individual growth curves. Part III, deals with anatomical and physiological age and growth in which among other measures Roentgenograms are used as criteria. Part IV is a historical survey of 911 investigations in the field, and Part V includes tabular summaries of all available data, comprising nearly five and a half million cases. Part VI is an annotated bibliography of the 911 articles and Part VII gives tables of English equivalents for the French metric system.

The monograph has no peer as a manual for advanced students and for teachers in this field. It is exhaustive in content and the technical manipulation of data is excellent. It contains important original contributions, especially in the form of continuous growth curves obtained by re-tests of the same children and the correlations of anatomical and with physiological, mental and emotional capacities.

A. I. G.

6. *An Elementary History of Education for Normal Schools.*²—This book by Dr. Finney is one of the new Modern Teachers Series, edited by Dr. W. C. Bagley. It is designed to give the prospective teacher some idea of the structure and purpose of American public education.

¹Baldwin, Bird T. *The Physical Growth of Children from Birth to Maturity*. University of Iowa Studies in Child Welfare. 1921. Vol. 1, No. 1. Pp. 411.

The book deals in an elementary way with 1. the Colonial period, 1667-1776; 2. Period of Nationalization, 1853-1861; 3. The Great Educational Awakening, 1853-1861; 4. The Transition Period, 1861-1890; and the Recent Period, 1890-1920. European influences are discussed for the most part by the inclusion of chapters on Rousseau, Pestalozzi, Herbart and Froebel. On the other hand, such topics as the English Poor Laws and the Apprenticeship system which had an important bearing on American education receive too little space. The recent period discusses (a) educational reorganization (b) enrichment of the curriculum and (c) educational theory and practice.

The book tends to be encyclopedic in that it discusses many important movements in a paragraph. References and suggestive problems at the end of the chapters would have been of distinct help to the normal school student, for which it is designed. However, as a brief survey, made concrete by the inclusion of numerous pictures, charts and diagrams, the American Public School should help to make beginning teachers intelligent as to the historical development of public education and should serve to stimulate them to work for the improvement of the American school system.

E. U. RUGG.

7. *A Book Describing An Elementary School Curriculum Based on Year-Long Series of Projects.*—That school people run to educational “movements” has been exemplified by the recent fervor over organizing teaching methods on the basis of so-called “projects.” The chief protagonist of the movement definitely *restricts the use of the term to method*. His followers, however, are now engaged in rebuilding the curriculum on the basis of it. Probably the most extreme instance of such an application in curriculum making is Miss Wells’ *A Project Curriculum*.¹ Dr. Wells has attempted to construct a curriculum in which *all* instruction of each of the first six grades is organized in one continuous year-long series of activities or “*projects*.”

Children were actually taught by this method under her supervision in a school in Trenton, N. J., in the first three grades. Her course for the fourth, fifth and sixth grades is still theoretical and

²Finney, R. L. *The American Public School*. New York: MacMillan. 1921. Pp. XIV+324.

¹Wells, Margaret E. *A Project Curriculum*. J. B. Lippincott Co., Philadelphia. 1921. Pp. XIII+338.

is sketched only in outline in her book. The first grade "project" is *Playing Families*; the second grade, *Playing Store*; the third grade, *Playing City*.

This pedagogical innovation is based upon the notion that each phase of school work shall be so far as possible a replica of a life activity. According to this theory, since people live in families the children shall be actually organized as families and the entire work of a school year shall be carried on as a series of things which an individual in a group or a family would do. Thus, children are assigned different roles in the family; the work of the year is thoroughly dramatized; "doll families" are made and dressed, thus providing the "motivation" (the "stimulating environment" of the free-educationists) for even the arithmetic! And most of these curriculum reformers would go so far as to say that such situations—such "life activities"—*must* be found through which *all* the skills, *all* the necessary information and *all* training in problem solving, and development of fundamental attitudes is to be developed. The protagonists of such a method do not compromise with those who demand definite and economical practice on socially worth-while skills. They say, with Miss Wells, that you get the skill without specific repetition, under the "intense motivation" of "life situations"!

We have insufficient review space for a detailed critical analysis of this book or this theory. An article or a monograph should be written upon it, showing of how much of current psychology these believers fail to make use. Suffice it to say here that Miss Wells reports in detail the theses and principles upon which her curriculum is based. She attempts to show outcomes but succeeds very imperfectly. She merely lists the facts, skills, habits, attitudes, and ideals which must have been employed by the children. How much skill? How well are the facts learned? What problem solving abilities have been developed? The answers to these questions neither we nor Miss Wells know for, astonishing as it may seem, she did not measure to find out!

This book may be regarded only as an interesting suggestion applying the motives of "free" education, of a certain form of "project method" to the making of the curriculum—an application which it is very doubtful indeed if the leaders themselves in these movements will accept.

H. O. R.

..9. *Empirical Studies in School Reading*,¹—Under this title one of the "Teachers College Contributions to Education" undertakes, among other things, to analyze and classify the study helps in four sets of Literary Readers used in grades IV to VIII. No attempt is made to report or analyze the literary content of these readers in the light of the valuable criteria assembled and made available in the opening pages of the study. The point of view is that of method. Considerable value attaches to the assembled authoritative quotations on the nature and purpose of literature and the aims various methods of studying it. The compilers of literary readers are not listed among those whose writings were canvassed. While prefatory statements in their compilations are in general agreement with the aims and purposes advocated by the literary authorities previously mentioned, the study helps, questions and directions to pupils are decidedly formal in nature. Those who advocate in their texts so much language training and other formal work have, no doubt, considered this training essential and have thought to improve such training by using material of high literary value. That literary appreciation is another matter, hardly attainable as a by-product of such instruction is clearly demonstrated by the verbatim reports of experimental lessons and the experimental evaluation of methods and devices.

While the methods of study outlined in the books analyzed did not vary substantially from the methods used by a random sampling of Chicago teachers, none of the teachers used the helps suggested in the readers nor did the pupils refer to them.

One hundred and thirty-one persons of long experience in school-work were asked to rank eighteen questions in the order of their merit as aids in the study of literature. These rankings emphasize the value of questions which draw on the pupil's related experiences and assist him to realize in imagination the experiences of the poet. Matters of fact and literary technique were ranked low. While these judgments contradict teaching and text book practice, they harmonize with the preponderance of expert opinion of writers on literature and literary study.

Of the two methods set up and used experimentally, and tested under controlled conditions, the one which emphasized the sym-

¹*Empirical Studies in School Reading*: James Fleming Hoscic, Ph. D., Teachers College, Columbia University Contributions to Education, No. 114. New York City. 1921. Pp. VIII—174.

pathetic approach and gave opportunity for picturing and imaging the experiences in the selection was found superior to the one which stressed technique, analysis and factual detail.

The conclusions reached in the study merit careful consideration by those who are interested in the improvement of textbooks, and by those who realize the psychological significance of the aesthetic experiences to which the study of literature may open the way.

L. Z.

III. ADDITIONAL PUBLICATIONS RECEIVED.*

A. MENTAL AND EDUCATIONAL TESTS.

SANTA ANNA (CAL.) PUBLIC SCHOOLS, DEPT. OF RESEARCH. *Four Years of Standard Tests and Measurements*. By Mary B. Henry, 1921, Paper. Pp. 27.

B. PUBLICATIONS IN THE GENERAL EDUCATIONAL FIELD.

BERKSON, I. B. *Theories of Americanization*. Teachers College Contributions to Education; No. 109. New York: Teachers College, Columbia University, 1921. Pp. VIII + 226.

HALL, G. STANLEY AND HIS STUDENTS. *Aspects of Child Life and Education*. New York: Appleton & Co. 1921. Pp. XV + 326. (Reprint.)

HOSIAC, J. F. *Sample Projects*. 506 W. 69th St., Chicago. 1920. Paper. Pr. 32.

POWERS, S. R. *A History of Teaching of Chemistry in the Secondary Schools of the United States Previous to 1850*. University of Minnesota, Minneapolis, Minn. 1920. Paper. Pp. 69. 50 cents.

REAVIS, G. H. *Factors Controlling Attendance in Rural Schools*. Teachers College Contributions to Education No. 108. New York: Teachers College, Columbia University. 1921. Pp. 69.

THORNDIKE, E. L. *The New Methods in Arithmetic*. Chicago and New York: Rand McNally Co., 1921. Pp. VIII + 260. To be reviewed in the October issue.

C. NEW SCHOOL TEXTBOOKS.

CORYELL, H. V. AND HOLMES, H. W. *Word Finder*. Yonkers (N. Y.) World Book Co., 1921. Pp. VIII + 150.

DUNN, ARTHUR W. *Community Civics and Rural Life*. Boston; D. C. Heath & Co. 1920. Pp. XII + 507.

FINCH, CHAS. E. *Everyday Civics*. New York: American Book Co. Pp. X + 326.

*Publications which are reviewed in this issue are not listed here.

THE JOURNAL OF EDUCATIONAL PSYCHOLOGY

Vol. XII

October, 1921

No. 7

THE READING PROBLEM IN ARITHMETIC

PAUL W. TERRY
University of Washington

DESCRIPTION OF THE INVESTIGATION

It is rapidly becoming a matter of general recognition among school people that a definite and distinctive problem in reading is to be found in each of the subjects of the elementary school program of studies. The subject of arithmetic, which offers for reading such characteristic materials as verbal problems and numerals which are not set in problems, is not exceptional in this regard. In view of this fact, teachers of arithmetic are confronted with the necessity of developing a specialized technique for the reading of arithmetical materials. Modern scientific studies in the psychology of reading have given but scant attention to the reading of numerals. Such interest in numerals, as has been manifested, has proceeded from the point of view of pure science rather than from that of practical educational applications. Little is known, therefore, of the methods employed by children in the gradual acquirement of the power of reading numerals. A simple approach to this general problem may be found in the study of the methods used by adults in this process. It is the purpose of this article to present a few of the more important results of such an investigation.¹

The subjects who served in the several studies of the investigation were all graduate students of the School of Education of the University of Chicago. In each study from four to ten subjects were used. With only two exceptions their experiences with numerals had been entirely normal for adults of similar training. The instructions which were

¹ For a complete discussion of the investigation see Terry, P. W.: *An Experimental Study of the Reading of Isolated Numerals and of Numerals in Arithmetic Problems*, Supp. Ed. Mons. No. 18, Dept. of Ed., Univ. of Chicago.

given the subjects were designed to secure normal procedures on their part. They were asked to entertain their usual problem-solving attitudes and to proceed at normal speed. The problems were to be read and solved. In every case, opportunity was given to the subject before he began his work to examine materials similar to those to be read or solved, and under similar conditions.

The materials which they read were of two different kinds. First, there were ordinary arithmetical problems which were so designed as to include the desired numbers of the numerals that were to be studied. Some of the problems included two numerals; others as many as four. The lengths of the numerals varied from one to seven digits and numerals of similar lengths only were placed in any one problem. Problems A and B will serve as illustrations of the problem type of material.

PROBLEM A

At 47 cents a dozen what will 2 dozen eggs cost?

PROBLEM B

If one telephone company uses 1,918,564 cross-bars during the year, and another company, in the same period, uses 617,453 cross-bars, how many more does the one use than the other?

In the second place, there were ordinary numerals which were placed in lines and isolated from each other and from any other context. These numerals were from one to seven digits in length and included a representative number of each length.

The methods which were followed in procuring the data were three in number. In the first place, the adult subjects of the investigation were asked to make introspective observations of their procedures with the numerals. This they were easily able to do after the brief periods of training which were given. In the second place, the introspective reports of the subjects were supplemented by direct observations of the readings which were made and recorded by the author of this report. The information gained by the use of these two methods served as a basis of interpretation of the data secured by the third and more objective method of photographing the movements of the eyes while reading. The eye-movement photographic apparatus was that which is described by C. T. Gray in his monograph, *Types of Reading Ability as Exhibited through Tests and Laboratory Experiments*.¹ For the purposes of this article it will be sufficient to say that by the use

¹ Supp. Ed. Mons. V. I No. 5 : 83-90.

of this apparatus records were obtained which could be so interpreted as to show what words or numerals (or digits of numerals) were the objects of the reader's attention at any fixation of the eye. At the same time accurate records of the duration of fixations were secured.

THE FIRST READING PHASE

The first important phenomenon, which came in evidence frequently during the course of the investigation, was the fact that two separate phases are distinguishable in the reading of arithmetical problems. These phases differ both in time and in purpose and may be designated as the *first reading* and *rereading* phases respectively. When both phases were found in the reading of a problem, the general procedure on the part of the subject was first to read the text through with more or less attention to the numerals in order "to get the sense" of the problem (first reading). When this was done his attention was directed for the second time to the numerals of the text with the intention of perceiving them accurately and completely (rereading).

During the first reading of a problem the numerals are read with widely varying degrees of attention. In some cases, subjects attended with care to the identity and place of each component digit of the numeral, at the same time noted its character as whole or decimal and gained an accurate notion of its magnitude. Such detailed perception of a numeral is designated as *whole first reading*. In many cases, on the other hand, numerals were not read in this detail. At times the numeral was merely recognized as a numeral. In other cases the item of digit length, or digit length and identity of the first digit only, were observed. The first two digits only and digit length were also observed in some readings, and various other items (such as the fact that one numeral was larger than another) were reported by subjects. In all of these instances the perception of the numeral is indistinct and lacking in detail. Such reading of numerals is called *partial first reading*.

Important conclusions as to the significance of partial reading can be drawn from a study of the frequency with which the subjects observed such details of the numerals as are listed above. In Table I is presented under five classifications the range of correct recall of the numerals which a group of subjects was able to report immediately after the first reading of a set of problems. No attempt has been made to separate the results of partial reading from those of whole reading. Conclusions will be drawn concerning only those ranges of recall which

were reported in a decisively preponderant number of cases. Such ranges of recall are obviously characteristic results of partial reading as well as of whole reading. The facts which stand out most strikingly from Table I are that almost without exception the numerals were at least partially read, and that in nearly every instance their digit length was noted. The identity of the first digit also was so frequently recalled as to be classified as an equally characteristic result of partial reading.

Such details of the numerals apparently are of the same value to the subject as the general conditions of the problem. In consequence, these preliminary perceptions of the numerals are obtained contemporaneously with recognition of the general conditions. The significance of partial first reading, therefore, appears to lie in the fact that it enables the subject to think about the problem without entering at first into the minute details of solution.

The number of pauses of the eye and the total time required for the reading of any longer numeral vary, according to the type of first reading which is used. When the numeral 1,918,564 was read in the partial fashion by Subject G only two pauses of 9 and 18 fiftieths of a second respectively were recorded. When the same numeral was read in detail by Subject B, however, five pauses of 8, 9, 20, 18 and 36 fiftieths of a second respectively were required. In the same manner, the number of digits, which are read during one pause of the eye in partial reading is decidedly greater than the number read during a pause of whole reading. In the former case for the most part from three to four digits are read per pause, whereas in the latter, one and two digits are more frequently included in one pause.

The percent of partial and of whole readings which a numeral received from a group of subjects was found to vary with its length, the number of other numerals in the problem, and with its position in the problem. Longer numerals varying in length from four to six digits were read partially in decidedly more than half of the cases whereas exactly the reverse is true for the shorter numerals of one and two digits. Significantly larger percents of numerals of a given length were read partially when as many as four numerals appeared in the context of a problem than when only two numerals of the same length appeared. The first longer numeral to appear in a problem in which several longer numerals were included, received a larger number of detailed readings than the other numerals.

Another factor which determines whether a numeral will be read partially or in detail is found in the attitudes of the individual subjects

TABLE I
Range of correct recall of numerals from first reading of problems

	1 and 2 digit numerals						3 to 7 digit numerals							All num- erals
Problems	A	B	F	ABF	C	D	E	CDE	A-F					
Numerals read in problems	43 2	5 45 8	10 9 3		246 1,754	1,000 1,276 91 718 2,981,534	617,453							
Total number of readings given each numeral by all subjects.	7 7	7 7 7 7	6 6 6 6	53	6 6	6 6 6 6	7	50						103
Range of correct recall of numerals:														
Complete.....	5 6	6 6 7 6 6 6	6 6 6	48	6 1	5 1 0 1	0	14						62
First two digits and digit length.....	5	6 ... 6 ...	6 ...	17	6 4	5 3 0 1	1	21						38
First digit and digit length.....	6 6 6 7 7 6 6 6	7 7 6 6 6	50 6 5 5	50	6 6 5 5	4 2 1 6	3	32						82
Digit length.....	6 6 6 7 7 6 6 6	7 7 6 6 6	50 6 6 5	50	6 6 6 6	6 4 3 7	7	44						94
Merely noticed.....	1 0 0 0 0 0 0 0	0 0 0 0 0 0 0	0 0 0 0	1	0 0 0 0	1 0 0 2	0	3						4

toward the numerals. The shorter numerals of one and two digits are read in detail almost invariably by all subjects. To the longer numerals, however, some of the subjects give partial readings with remarkable consistency; whereas other subjects persistently read the same numerals in detail. The former may thus be classified as *partial first readers* of numerals and the latter as *whole first readers*. A clear illustration of the difference in the readings given to a set of numerals by partial first readers and by whole first readers is found in a comparison of the records of subjects G and H, each of whom read a certain set of seven problems. There were nineteen numerals in the set. Six of the numerals were one or two digits in length and thirteen were from three to seven digits in length. Subject G, a marked partial first reader, gave partial readings to fourteen of the numerals, while Subject H, a consistent whole first reader, read all of the numerals in detail.

Attention has been called in paragraphs above to the fact that numerals in problems are attacked by partial first readers with attitudes which differ distinctly from the attitudes of whole readers. The very fact that attitudes towards the numerals *do* vary so widely suggests that the numerals stand out from the other contextual elements of problems in some distinct way. An experiment was arranged, therefore, with a view to determining in what way the numerals and words, which constitute the text of an arithmetical problem, differ in their demands upon the attention of readers. The records from this experiment show that the numerals of problems make decidedly greater demands upon attention than the accompanying words. Less than half as many digits as letters is perceived during one pause of the eye. The average duration of pauses on numerals was found to be approximately 40 per cent greater than for pauses on words. The per cent of regressive pauses in the total number of pauses, is much greater for numerals than for words. In respect to each of these three items the numerals make greater exactions of the readers, and this is true in spite of the fact that many of the numerals were read in the cursory partial fashion. The explanation of this contrast probably lies in the differences between numerals and words in point of construction. The letters in words appear and reappear in the same regular combinations which become familiar in the earlier years of schooling and are read as wholes. The digits in numerals, however, appear in constantly changing combinations. Each individual digit is significant in itself and all of the digits must be perceived before the numeral is accurately read. It is obvious, therefore, that the

mind is more occupied with processes of analysis and combination of component digits when numerals are being read than with similar processes when words are read.

COMPARISON OF PARTIAL AND WHOLE FIRST READING

The above discussion provides ample evidence for the conclusion that the numerals are a unique feature of the reading situation which is involved in the reading of arithmetical problems. Since two distinctly different methods have been developed for reading the numerals, the question necessarily arises: which is the more efficient? The answer depends in large measure upon the subsequent question, which of the two methods is more economical of the reader's time? Attention should be called at this point to the fact that the type of activity which was displayed by the subject after the first reading in his efforts to solve a problem did not appear to depend upon which method of attending to the numerals was followed during the first reading. Under these conditions, the first reading may be treated as a relatively independent phase of the solving of a problem and likewise the question of the comparative efficiency of the two methods of first reading may be considered independently of activities subsequent to the first reading.

TABLE II

Comparison between partial and whole first readers of numerals in respect to rates of reading

(Time unit = $\frac{1}{50}$ seconds)

Subjects	Partial readers			Whole readers		
	G	M	W	B	H	Hb
Total time required to read the numerals of the five problems ¹	195.0	304.0	245.0	480.0	359.0	337.0
Total time required to read the words of the five problems ¹	708.0	800.0	925.0	560.0	1,092.0	923.0
Average reading time per line with ordinary prose ²	52.5	44.9	72.2	80.8	75.0	

¹ A set of five ordinary arithmetical problems including 12 numerals and 111 words, which were read before the photographic apparatus.

² An ordinary expository prose passage of 10 lines from Judd's *Psychology of High School Subjects*, p. 190.

With this question in view data have been arranged in Table II from the readings of five ordinary arithmetical problems and a selection of ordinary prose by six subjects. Three of the subjects used the partial method and three the whole method of reading the numerals. By inspection of the first row of the table it is readily seen that the three partial readers, (G, M and W) read the numerals more rapidly than the three whole readers. The case for the words of the problems is not so clear, although the second and third fastest subjects were the partial readers, G, and M, and the slowest subject was the whole reader H. With the selection of ordinary expository prose, the fastest reading was clearly done by the subjects who used the partial method with the numerals in the problems.

Partial first reading of numerals is essentially a method by which a smaller number of pauses is used than in reading in detail. The greater speed of the partial readers with the ordinary prose selection also was due primarily to the use of a smaller number of pauses. Apparently, the virtue of partial reading of numerals and the greater rapidity of the partial readers with problems and with other materials as well is due to the fact that the same quantity of reading material is read with fewer pauses of the eye. In point of time it is the more economical method and therefore as far as the materials and subjects of this investigation afford a basis for conclusion, partial reading appears to be the more efficient of the two methods.

REREADING

The second phase of the reading of a problem—the rereading—, follows immediately upon the completion of the first reading. The rereading is concerned almost invariably with numerals only, although instances were recorded when words also were reread. Numerals of from one to seven digits can be read by one act of rereading. Numerals one to four digits in length were invariably reread at one reading. Six and seven digit numerals in several instances required two acts of rereading. In such cases the first several digits were read as a group and copied on paper, whereupon immediately the remaining digits were read and copied. Numerals of greater digit lengths required greater amounts of time for rereading than did shorter numerals. The total time required for the rereading of a problem, as would be expected, was decidedly less than that required for the first reading.

Two types of rereading were found which are clearly distinguished by differences in both function and procedure. The first type, which

may be described as simple rereading, has for its function the securing of further information concerning a numeral before a decision has been reached as to how to proceed with solving the problem. Only one of the numerals of a problem is selected for this type of rereading, for the most part. The object of the reader is very specific. He looks for such detailed items as the number of digits, the identity of certain digits and the location of the numeral in the line of print. The second type of rereading has for its function, apparently, a careful inspection of the numerals as a step preliminary to copying them. After each case of rereading for copying, the records show that the subject immediately proceeded to copy the numeral on the paper upon which computation was to take place. It thus appears that whenever this type of rereading was used the subject had already advanced with the solving of the problem to the point of deciding that the numerals should be copied.

REREADING PROCEDURES DURING COMPUTATION

It was found impossible to interpret precisely the photographic records of the eye-movements of subjects while they were engaged in computing from copied numerals. Several of the subjects, however, exhibited another method of computing, the records of which were more susceptible of accurate interpretation. In this case the procedure of the subjects was to treat directly with the numerals as they appeared in the context of the printed problem and to compute "mentally" without copying the numerals. This method, which is called direct computation, was found practicable under widely varying conditions. It was used by both partial and whole first readers, with both longer and shorter numerals, and after partial as well as whole first reading of numerals.

It is obvious that certain ways of reading and solving arithmetical problems which have been described in this investigation are more economical than others. When the method of direct computation is used, not only is the laborious step of copying the numerals saved but also the additional step of rereading, for only in rare instances were numerals reread which were not to be copied. It is important to recall at this point that direct computation without rereading was practiced even when the numerals had been read only partially during the first reading. The records furnish conclusive testimony to the striking fact that only such details of numerals as can be perceived with the rapid and cursory attention which is given during partial reading, are re-

quired as a basis for actual computation. In cases of this kind the numerals were not read in complete detail until they were attacked in the process of computation itself. The conclusion may be drawn, therefore, in so far as this investigation is concerned that the procedure of direct computation based upon a partial first reading of the numerals is the most economical way of reading and solving arithmetical problems.

During the process of "mental" computation direct from the printed lines, the numerals which appeared in a problem were not treated with equal attention. This fact was observed in the records of only those problems which contained exactly two numerals. In several such instances one numeral was taken as the "base of operations." When a numeral was so used its digits served as the starting point of the computation and the digits of the second numeral were related to it. More pauses of the eye and pauses of greater average duration were located on the digits of the "base" numeral than upon the digits of the second numeral.

The numeral of greater digit length was selected almost without exception as the "base of operations." Such a selection is in keeping with the common practice of placing the numeral of greater magnitude first in the order of computation and relating the smaller numeral to it. The larger number of pauses of the eye on the "base" numeral is probably due to the fact that this numeral presents one digit more for reading than the second numeral, and to the further fact that computation both begins and ends on the "base" numeral. The fact that the eye lingered for longer pauses on the "base" numeral suggests that a peculiar quality of work was done during the pauses on this numeral. It is the author's opinion that when the eyes of the subject were engaged with the digits of the second numeral, the work done was in the main simply that of recognizing the digits which were being read. On the other hand, the work which was accomplished during the pauses on the "base" numeral included not only recognition of the digits which were being read but also processes of a more definitely arithmetical quality. For such work it appears reasonably certain that additional time would be required.

ISOLATED NUMERALS

The isolated numerals were arranged for reading in two different ways. In one section of the investigation they were placed in columns, one numeral only to a line. In a second section they were

placed in regular printed lines and at such distance within the line from each other that the reading of one numeral did not interfere with the reading of any other numeral. The space between the lines of numerals in both sections of the investigation was arranged with a view to preventing the subjects from being occupied with numerals in more than one line at a time. The instructions to subjects called for a low and easy articulation of the numerals as they were being read. The provision for articulation enabled the author to hear and record the grouping of the digits of the numerals and the numerical language which was used. In the second section of the investigation photographic records were made of the movements of the eyes of the subjects while they read the numerals. The data which were procured by this method supplemented those which were obtained by recording the subject's articulations.

One of the most striking features of the reading of numerals is the fact that the subjects habitually divided the numerals into digit groups. When this is done certain successive digits are so closely associated with each other in the reading as to form units of reading, which units are at the same time distinguished from other similar units. The digits which constitute a group are bound together by being pronounced in quick succession as one series. The pronunciations of the several digit groups are separated from each other by time intervals distinctly longer than the intervals which separate the pronunciations of the individual digits. Three different sizes of groups were clearly distinguished in the readings, namely, those that were made up of one, two and three digits respectively. The one digit groups appeared more frequently in the one, three and seven digit numerals. The two and three digit groups appeared in the readings of numerals of all the greater digit lengths.

It became apparent early in the course of the investigation that the digits of numerals of the same lengths were being grouped in much the same manner by all of the subjects. The outstanding fact is that the digits of numerals of any particular length are divided into a certain number of groups made up of a certain number of digits, which groups stand in a certain order of succession. Such an arrangement of the digits of a numeral is designated as a main group pattern. The one and two digit numerals were each read as single groups of one and two digits respectively. The first variation from one main group pattern as representative of the reading of numerals of the same length, occurs in the three digit numerals which exhibit two patterns. The

four digit numerals appear almost invariably in a pattern of two groups of two digits each. The five digit numerals show in a preponderant number of cases a pattern of two groups of two and three digits respectively, and the dominant pattern of the six digit numerals is that of two groups of three digits each, while a three group pattern of one, three and three digits respectively was used for seven digit numerals.

Approximately one-half of the five to seven digit isolated numerals in one section of the investigation was not written with the usual comma punctuation. Opportunity was afforded in this way of studying the effect of punctuation on the grouping of the digits of the longer numerals. Punctuation apparently operates in decisive fashion to increase the number of three digit groups used, and conversely to decrease the number of groups of two digits. Fewer main group patterns appear in the readings of the punctuated numerals. By thus regenerating the main group patterns, and in encouraging the use of the larger group of three digits, the employment of punctuation appears to give greater facility and speed to the reading of numerals.

When a detailed examination is made of the eye movements with which the numerals were read, it appears that two distinct types of pauses are represented. Pauses of the first type, which may be called *strictly reading pauses*, were probably used in recognizing the identity of the digits of the numerals and the relations between the digits. Such pauses are invariably located on the numerals and their durations are approximately equal to, or greater than the average duration of the pauses of the subject whose records are under consideration. A preponderant number of the pauses of any subject are of this first type. Pauses of the second type, which are called *guiding pauses*, were probably used in locating the first digits or the last digits of the numerals. They are found on the initial or final digits and more frequently on numerals of greater lengths. Some of these pauses appear on the lines between two numerals. Their duration is very brief as compared with that of the other type of pauses.

The number of pauses and the total time, which is required to read a numeral, depend on the digit length of the numeral. The average total reading time per numeral increases steadily from the 21.45 fiftieths seconds average for the one-digit numerals to the 104.54 fiftieths seconds average for numerals of seven digits. Likewise the average number of pauses per numeral increases steadily. Numerals of the same length exhibit a notable consistency in the number of pauses with which they were read. The average duration of the

pauses, on the other hand, does not depend upon the length of the numerals in the same consistent fashion.

The subjects did not uniformly follow one single style of attack as they read those isolated numerals which were arranged several to a line. Two radically different modes of perceiving the numerals were displayed. By one method a relatively large number of pauses of relatively short average duration were used. A significant proportion of these pauses were of the guiding type. By the other method relatively few pauses of relatively long average duration were employed. The shortest total reading times for the whole set of isolated numerals were found in the records of the two subjects who used the many-short-pauses method of attack.

AN EXPERIMENTAL AND STATISTICAL STUDY OF READING AND READING TESTS

(Continued)

ARTHUR I. GATES

Teachers College, Columbia University

THE THORNDIKE AND THORNDIKE-McCALL SCALE FOR THE UNDER- STANDING OF SENTENCES

Thorndike and McCall have devised 10 new forms, identical in principle with the test reported by Thorndike in 1915.¹ In the latest editions the test consists of 11 paragraphs, increasing in difficulty, which are to be read for the purpose of answering three or four questions which follow each paragraph. The number of questions correctly answered in 30 minutes may be transmuted into a scale score, in terms of which norms for ages and grades are given. Adjustment to the work in the test is secured by a fore-exercise. This test is of the sort frequently called a "power" test: it aims to discover how difficult material the subject can comprehend.

The original Thorndike Alpha and forms 1 and 2 of the McCall revision were given to all grades. Forms 1, 2, 3, 4 and 5 of the latter were given on successive days for experimental purposes to grades IV and VI.²

The Difficulty of Forms 1, 2, 3, 4 and 5, Thorndike-McCall and the Consistency of Performance in the Several Forms

Forms 1 to 5 which were given on successive days to the pupils of grades IV and VI appear in Table III.

TABLE III

	Forms	1	2	3	4	5	Average	S.D.
Grade 4.	Mean	49.7	50.3	48.0	45.0	50.3	48.6	1.9
	S.D.	3.4	4.6	5.2	4.5	5.5	3.3	
Grade 6.	Mean	56.5	56.7	61.1	58.4	59.6	58.3	1.8
	S.D.	4.0	6.7	5.9	5.6	5.6	4.7	

The several forms appear to be of approximately equal difficulty. The difference between scores 56 and 61 is the difference between

¹ *Teachers College Record*, November, 1915 and January, 1916.

² We are indebted to Miss Lulu Ailes and Miss Bess Young, teachers of grade IV and VI respectively for giving these tests.

answering 27 and 29 questions, which is not excessive. The order from easy to difficult forms differs for the two grades, the correlation between the two orders being in fact—0.10. This may well be expected, for in this test the upper limits of comprehension are at different levels. This fact makes it very difficult in tests of this type to secure consistency in performances for the individual child even if mean scores for a grade are equal, for the reason that the material in the vital area—at the mean limit of comprehension—(questions 22–24 for our grade IV) may vary in difficulty for individuals because of the particular content. * The correlations of single tests are likely, then, to vary widely.

Table IV gives the correlations between each single test and the composite of all five tests, as well as a few correlations between the different forms.

TABLE IV

Form	2		3		4		5		Composite	
Grade	IV	VI	IV	VI	IV	VI	IV	VI	IV	VI
Form 1.....	0.25	0.41	0.52	0.70
Form 2.....	0.72	0.56	0.77	0.75
Form 3.....	0.36	0.63	0.77	0.82
Form 4.....	0.06	0.55	0.45	0.79
Form 5.....	0.73	0.83
Mean.....	0.65	0.78

As expected, the inter-correlations of single tests vary widely but the correlations with the composite are, of course, much higher. All of these correlations are, however, very misleading unless the composition of our groups is considered; they represent a very narrow selection of the school universe in terms of performance in this test. The mean score for grade IV is 48.6 with an S.D. of 3.3: for grade VI, the mean is 58.3, S.D. 4.7. The S.D. for grade VI represents but $1\frac{1}{2}$ answers from a mean of 28. The correlations for grade VI, it will be noticed, are larger in the mean than those for grade IV for the reason, almost certainly, that the S.D. is larger. With a random sampling of pupils, the correlations would be even larger.

The consistency of group performance appears more clearly when the scale score is converted into "Questions answered." For example, the S.D. of the mean scores on different forms from the composite is 1.9 for grade IV (see Table III, last column). This means a S.D. of approximately three quarters of a question on a basis of 23, which is the mean for the grade. For grade VI, the S.D. is about $\frac{2}{3}$ of a question on the basis of 28. This indicates a rather high degree of consistency in performance for the class as a whole in terms of the test units. The units, however, are really very coarse, so that when we convert them into grade norms, the variability of the class means on the several forms becomes so great that the material cannot be used instructively for measuring monthly improvement, even for groups.

In Table V the scores of Table III have been converted into grade norms, using the tables provided by the authors of the test. The S.D.'s are 0.32 and 0.75, of a school year.

TABLE V
Showing the grade norm equivalent of the mean scores yielded by each form

Forms	1	2	3	4	5	Com- posite (mean)	S.D.
Grade IV.....	6.2	6.4	6 0	5.5	6.4	6.2	0.32
Grade VI.....	7.5	7.7	9 2	8.0	8.5	8.0	0.75

An inspection of the grade status yielded by the individual tests reveals wide variations. The variability for individuals is, of course, very much larger. It is quite clear that if a precise measure of an individual is required several forms must be given or new forms, including more numerous and much smaller steps, must be devised.

The correlations of the composite scores of 2 or more tests with the composite of all 5 follow:

	2 with 5	3 with 5	4 with 5
Grade IV.....	0.70	0.85	0.96
Grade VI.....	0.75	0.88	0.97

Perhaps the following table showing the deviations of 1, 2 or more tests composites from the composite of all 5 is more illuminating:

TABLE VI
Deviations from the composite of 5 tests

	Grade IV		Grade VI	
	In scale score	In grade years	Scale score	Grade years
Composite of 4 tests.....	0.35	0.04	0.1	0.033
3 tests.....	0.70	0.112	0.2	0.066
2 tests.....	1.40	0.224	1.7	0.560
1 test.....	1.90	0.314	1.8	0.590

This table shows that if we give but one test, the class mean (grade IV) will deviate 0.31 of a school year's progress from the true score. Calling 9 months a school year, if we give two tests and combine them, the class score will deviate 0.22 of a year; for 3 tests about 1 month, for 4 tests about $\frac{1}{5}$ of a year. This means that if we wish to measure monthly progress, we will need to secure a composite score of 3 or 4 tests, requiring $1\frac{1}{2}$ to 2 hours time. To measure adequately the monthly progress of each individual will require at least 3 tests.

The data illustrate in an interesting way what seems to be a fact that even the more carefully constructed of our educational tests are insufficiently refined for exact individual examination, in short periods of time. They are extremely useful however, since, inexact as they are, they are much more accurate than any information otherwise obtainable and since exact measures can be secured if enough time is given to this investigation.

The Effects of Practice.—What we know of the effects of practice makes it certain that many functions with repeated testing will show large improvement. This improvement is specific in character, and we are not, in such cases, warranted in assuming that "general reading ability," for example, has shown a corresponding development. It is imperative that each test be carefully examined for purposes of measuring the specific improvement.

A glance at Table III will show that the effects of two hours or more of practice on Thorndike-McCall is very small. The improvement is obscured by differences in difficulty of the tests, but the irregularity of scores shows that improvement cannot be great.

Comparison of Thorndike Alpha 2 and the Thorndike-McCall.—The Thorndike-McCall is essentially the same as the Alpha 2 though

differing mechanically in the method of computing scale scores and in the groups from which the norms were secured. The correlations of Alpha 2 with any form of Thorndike-McCall is about the same as the inter-correlations among forms of the latter. The grade norms yielded by the two were found to vary but little, in the lower grades but very greatly in the upper grades. Table VII shows the data. The grade norms were based on tests given Feb. 1.

TABLE VII

Average inter-correlations of Thorndike-McCall		Correlations of Thorndike-McCall with alpha 2	Mean grade norm alpha 2 ¹	Mean grade norm Thorndike-McCall
grade 3	0.55	0.72	3.25 ¹	3.2
4	0.57	0.55	5.6	5.4
5	0.58	0.40	6.15	6.6
6	0.57	0.33	7.30	7.2
7	0.60	0.52	8.50	11.2
8	0.59	0.60	8.70	11.0
mean	0.57	0.53		
S.D.	0.02	0.11		

Correlations of Thorndike-McCall with other Criteria.—A survey of Table VIII reveals certain facts about the Thorndike-McCall test; (1) It yields a correlation of 0.50 with almost any other single test of comprehension except the Brown, and a somewhat smaller coefficient with single measures of rate. It yields correlations of 0.50 with vocabulary tests and 0.57 with an oral reading test. The correlation with the corrected composite is 0.73 as compared to 0.52 with the composite of speed. The latter correlation is significant however, since this test more than any other makes no pretense at measuring speed. The mean correlation with the composite of group intelligence tests is 0.69 showing that in all likelihood the ability to "understand sentences" is considerably involved in most of our group tests.

The correlations with the Stanford Mental Age are interesting, in particular the fact that the coefficients become rapidly higher as we advance from grade to grade. There is a temptation to jump to the conclusion that reading bears a high relation to general intelligence

¹ The figure 3.25 means one-fourth of the way through grade III; 3.0 would mean at the beginning, 3.5, midyear, etc.

only when it reaches the upper limits of mechanical perfection, usually assumed for Grade V (or better Grade VI) but there are other possible explanations. If this hypothesis were correct, it would mean that our conventional group tests, depending considerably on reading, are inadequate in the lower grades. On the other hand this is not equivalent to saying that reading for a third grade child is a very variable performance (the inter-correlations among the reading tests for Grade III are, in fact, as high as for any other grade) but merely that the reading performance, though stable, is not highly correlated with general intelligence, if the Stanford Mental Age is an adequate criterion. This matter needs intensive investigation. It may be that we should inform teachers that the M.A. is a real measure of intelligence, but that it will not predict exactly a pupil's achievement in the lower grades. We expect to investigate the correlations with performance in other school subjects in a later paper.

The study of certain cases of difficulty in reading has yielded information concerning the functions measured by the Thorndike-McCall tests which has been obscured in the correlations.

Case S.S.¹—A girl in Grade VIII has an IQ known to be above 125. She is an excessively slow reader. Her difficulty is functional, *i.e.*, there was found no organic defect of any kind in any of her bodily or nervous mechanisms. Her oral reading is poor for Grade III. Her speed in silent reading measured by speed tests listed above, Gray's Silent Reading tests and informal tests, was found to be approximately one word per second. The Thorndike-McCall yielded a score of 65, compared to the mean of 63.4 for her grade. Score 65, in McCall's norms is that of a pupil half through the Grade XI of schools at large. We are convinced that this is an adequate measure of her "comprehension," excluding speed. No other test of comprehension gave her a high score. On the Courtis comprehension she scored 24, the next lowest being 35, the mean 55.6 for the class. Monroe's comprehension score was 15 compared to a class mean of 38; the Burgess gave her a score of 8, the class mean being 67.8 and so on. Several other cases yielded similar results which are, to our mind, rather convincing evidence of the diagnostic value of the Thorndike-McCall in critical cases. It probably is the only test measuring a certain type of "power" in comprehension, unaffected by the mechanical factors of reading. From our data it appears that the test is but little subject

¹ This case and others will be considered in detail in a later paper.

to improvement through specific practice, and there is consequently doubt as to whether it ever, except at beginning stages of learning to read, yields a measure of the amount of effectiveness of instruction. It has frequently been noted that schools in which little or no effort is made to teach reading, make a good showing on this test, whereas they may do badly in instructional material such as spelling and arithmetic.

It has also been said that this test is simply another measure of intelligence. It probably is a measure of one sort of "verbal intelligence" and is, on that account, one of the most useful of our tests. There is a great need for further investigation for purposes of discovering for certain whether the function represented in this test is one that will yield to practice, or whether it is one which develops primarily as a result of inner growth.

THE BURGESS PICTURE SUPPLEMENT TEST

The Burgess tests consist of a series of 20 pictures with a paragraph of about 55 words following each. The paragraph includes instructions to complete the picture by drawings or by writing words. The paragraphs are said to be of equal difficulty of vocabulary, phraseology and thought. The score is the number of paragraphs the directions of which are properly fulfilled in 5 minutes. Several forms of the test are now available.¹

In most respects the experimental, statistical and reflective work behind the Burgess tests is admirable. Since it appeared to be a test of great promise, certain experiments were devised to analyse further some of its features. Form I was given to all grades in the usual way, and about 7 weeks later the pupils of grades III, IV and V were tested individually with the same form. It was clear that no pupil had sufficient recollection of the material to influence the score. Each child went through the first twelve paragraphs and by the use of two stop watches the writer who conducted the tests, was able to measure (1) the time spent in actual reading and (2) the time spent in writing or drawing the answer for each paragraph.

The Reliability of the Burgess Scale.—The correlations obtained from the ratings of the two trials were:

¹ Burgess, May Ayres. *The Measurement of Silent Reading*. Russell Sage Foundation, Educ. Monograph, 1921. Pp. 163.

	Coefficient of correlation	Coefficient of reliability
For grade III.....	0.62	0.77
For grade IV.....	0.59	0.74
For grade V.....	0.66	0.79
Mean.....	0.62	0.766

The mean coefficient of reliability is slightly lower than that of 0.83 obtained by Mrs. Burgess on grades II to VI at the Lincoln School, a difference to be accounted for, very likely, by our smaller S.D's.

Mrs. Burgess, in her monograph, quite rightfully questions certain interpretations of coefficients of reliability. She says they give us "more information about the children than they do with regard to the test" and their test scores "may vary widely from day to day and still be actual true measures of ability on each occasion. Under such conditions the fact that the scores vary from trial to trial does not reflect any inaccuracy or inadequacy of the test." It may not, in some cases, but usually wide variability does indicate, in educational instruments, an inadequacy of the test. The scale, to meet present day demands, must either be improved or lengthened or both, so as to give consistent results. One factor which we can point out in passing is the coarseness of the units involved in this test. In grade IV, for example, nearly half a minute on the average is devoted to each paragraph. No credit can be earned unless the paragraph is completed. In grade III, subjects spend more than a minute on paragraph 8, a few spent nearly two minutes. Many children of really unequal ability thus earn the same score.

Another possible cause of apparent inconsistency in performance is the rather rough method of scoring. No credit is given unless the paragraph is correctly supplemented. We have found that experts in grading are frequently uncertain whether the markings "exactly follow directions." That others have had the same experience is stated by Whipple; "Those who are using the new Burgess Silent Reading Scale are raising many questions concerning the scoring of doubtful performances."¹ The list of rulings given by Whipple should have accompanied the original tests. When the penalty for an

¹ *Bulletin of the Bureau of Tests and Measurements*. University of Michigan, No. 18, 1921.

erroneous solution is as heavy as it is in this the test, the criteria, for grading should be very carefully standardized.

Mrs. Burgess in her monograph has specially insisted upon the observance of one of the commonly accepted canons of scientific methodology, namely the control of all variables save one—"Law of the single variable." She has endeavored more seriously than many other workers in this field to put this law into effect. "While the scale was being developed every endeavor was made to construct the paragraphs so that they should be of equal difficulty as reading material, of equal difficulty with respect to the instructions they contain, and of substantially equal requirements in the time necessary to read the paragraph and make the mark which supplements the picture." (p. 107). So far as empirical work is concerned this was done by giving the paragraphs in groups of 20, in one of two orders, to many children, counting the number of times that the instructions were successfully fulfilled. In the case of Form I, the paragraphs are of satisfactory equality with respect to this criterion. In spite of the unusual care taken to secure equality among the variables it appeared, when the test was first given at Scarborough, that the paragraphs were unequal in two respects which influenced upon the score, namely, the time required to read and the time required to draw the supplement. It also appeared that in many cases nearly as much time was spent in drawing as in reading, and if so, unless the correlation between reading and drawing was approximately $+1.00$, the test was subject to a very serious defect. The individual examinations were conducted to determine the facts.

Only those paragraphs which were correctly solved were considered. Table IX shows that in terms of the time required to complete the paragraphs, they are not of equal difficulty. Paragraph 8 requires more than twice as much time as 1 or 3 although, as Mrs. Burgess' data show it is correctly solved just as frequently, namely, in about 9 cases out of 10 attempts made by school children in general. The S.D. from the mean total times are large for all grades. Opportunity is afforded for wide variability of performance in individual cases and there is reason to fear that additional forms of the tests, if standardized by the single criterion employed might yield quite different scores. Additional forms were not in print at the time our experiment was conducted, but a recent study¹ shows that such is the case, Form 2 yielding a score about 8 per cent higher than Form 1.

¹ H. C. Daley: *Journal of Educational Research*, June, 1921. Pp. 71-72.

Table IX shows that the inequality of time required for the several paragraphs is due more largely to time spent in drawing than to time spent in reading. The S.D.s for reading are not unsatisfactory,

TABLE IX

Para- graph	Grade 3			Grade 4			Grade 5		
	A Mean total time	B Mean time to read	C Mean time to draw	A Mean total time	B Mean time to read	C Mean time to draw	A Mean Total time	B Mean time to read	C Mean time to draw
1	29.7	25.5	4.2	21.0	16.2	4.8	19.5	17.3	2.2
2	39.0	23.4	15.6	22.5	15.3	7.2	27.1	14.3	2.8
3	33.7	30.0	3.7	20.6	16.0	4.6	15.9	12.3	3.6
4	40.2	29.7	10.5	26.8	17.6	9.2	24.1	16.6	7.5
5	37.1	31.1	6.0	23.3	17.6	5.6	21.5	16.9	4.6
6	46.0	42.0	4.0	27.1	23.8	3.3	23.6	19.9	3.7
7	40.7	32.4	8.3	25.1	17.7	7.4	22.2	16.2	6.0
8	61.4	34.7	26.7	42.2	18.7	23.5	44.5	20.5	24.0
9	45.0	37.0	8.0	29.5	22.1	7.4	30.0	23.2	6.8
10	38.5	32.8	5.7	22.1	15.9	6.2	22.4	17.3	5.1
11	45.0	38.0	7.0	26.0	20.0	6.0	24.0	18.0	6.0
12	51.2	45.7	5.5	27.9	23.1	4.8	28.0	22.0	6.0
Mean	42.3	33.6	8.8	26.2	18.6	7.5	25.2	17.8	6.5
A.D.	7.8	5.6	5.8	5.5	3.0	5.0	7.4	3.1	5.5
Percentage of total timespent in drawing, per cent.....			21	29	26

although there is, to be sure, room for improvement. A more serious matter is the fact that one quarter of the total time is spent in drawing, and that the time varies greatly from paragraph to paragraph. For example, the range is from 3.3 to 23.5" for grade IV with a mean of 7.5" and S.D. of 5.0". The time for drawing likewise varies greatly from individual to individual. In drawing the three feathers to complete Paragraph 4, for example, some rapidly draw three dashes or

ovals, while others draw the details of a feather with great care. Time thus spent obviously has nothing to do with reading and the variability in drawing time will seriously reduce the validity of the test unless there should be a very high correlation between speed of drawing and reading ability. On *a priori* grounds we would not expect it. It should be noted, in passing, that our children were specially warned to waste no time in drawing and also that they have had unusual amounts of practice in taking all kinds of group tests.

Column A in Table IX gives the score which the test actually utilized, *i.e.*, mean total *times* to read and draw. Column B gives the mean time spent in reading. Correlating column A and B, it is found that $r = 0.52, 0.38$ and 0.18 for grades III, IV and V respectively. Since the frequency of making a correct solution is constant for the several paragraphs, these coefficients show that the scores actually utilized by the test parallel rather poorly the scores which represent reading time precisely. This is merely another way of showing that the validity of the test will depend in part on the correlation between speed of drawing and speed of reading.

Table X shows the correlations obtained from the speeds displayed by the subjects in reading, drawing, and the total of the two. Fortunately for the test, positive correlations of $0.16, 0.24$ and 0.23 were obtained for speed of reading with speed of drawing, which accounts in some degree for the coefficients of $0.69, 0.93$ and 0.85 between "total time" and "reading time." It is certain, however, that the large amount and inequality of the drawing time makes this test less useful than it might otherwise have been. Columns 4, 5 and 6, Table X

TABLE X
Showing correlations among several scores in the Burgess test

		2 Speed of reading	3 Speed of drawing	4 Speed on Para- graph 1	5 Speed on para- graph 6	6 Speed of para- graph 8
1. Speed of reading and drawing.....	Grade III	0.69	0.33	0.65	0.42	0.21
	IV	0.93	0.57	0.71	0.67	0.47
	V	0.85	0.48	0.68	0.71	0.38
2. Speed of reading..	Grade III	0.16	0.73	0.53	0.16
	IV	0.24	0.66	0.63	0.21
	V	0.23	0.63	0.52	0.29

show that paragraphs 1 and 6, in which time for drawing was little, yield a much higher correlation with the total score for reading than does paragraph 8 on which the extreme amount of time is devoted to drawing.

The results of the study of the Burgess test may be summarized as follows:

1. The units are rather coarse. This is not a valid objection if sufficient time is allowed. It is a more serious matter, of course, in the lower grades, where from 5 to 7 paragraphs only are read by the median child in the time allowed—5 minutes.

2. More specific directions for scoring should have been provided.

3. Too much time is devoted to drawing—about one quarter.

4. The various paragraphs, while equal in difficulty as regards success in supplementing the pictures, are not of equal difficulty on the basis of time required to complete the reading and drawing and are therefore not of equal value as measures of reading.

The general idea upon which the Burgess test is based is excellent. If the time for drawing were reduced to a minimum, and other changes as suggested above made, it would, in our opinion, be vastly superior to any existing test of reading rate. In its present form it is very useful, as will be indicated by the correlations which follow:

Correlations of the Burgess with Other Criteria.—Table XI gives the correlations of the Burgess test with the composites of rate and comprehension and the several separate tests. On the whole, the correlation with the composites are high, higher than those yielded by any other test and equally high with rate and comprehension; 0.82 and 0.80 respectively. Performance on the Burgess is relatively consistent from grade to grade, the S.D.'s being small relative to those shown by other tests (see Table I). Allowing for differences in the S.D.'s for grade performance, the test seems to be about equally useful in all grades with a possible exception of grade VIII.

The correlations with the Stanford-Binet are interesting. They increase from nearly zero for grade III to .56 for grade VI. The same relation was revealed by the Thorndike-McCall test. If these data represent the real relation of the sort of intelligence measured by the Stanford-Binet and reading ability, the grade differences will be of marked import.

Correlations with the composite of group intelligence tests are much higher and show no real variation as we pass from grade to grade. Reading ability is demanded in most of the group tests, and the

TABLE XI
Showing the Correlation of the Burgess Reading Test with—

Grade	(1) Stan- ford mental age	(2) Comp. group intell.	(3) Curtis rate	(4) Curtis comp.	(5) Brown rate	(6) Brown comp.	(7) Monroe rate	(8) Monroe comp.	(9) Thorn- dike McCall	(10) Direc- tions	(11) Special vocab.	(12) Holley vocab.	(13) Gray's oral	(14) Comp. comp.	(15) Correct comp. comp.	(16) Comp. rate
III	0.10	0.53	0.65	0.72	0.68	0.07	0.61	0.70	0.79	0.76	0.69	0.82	0.74	0.78	0.76	0.87
IV	0.13	0.55	0.69	0.38	0.60	0.01	0.70	0.66	0.34	0.80	0.63	0.51	0.58	0.82	0.84	0.94
V	0.53	0.54	0.37	0.24	0.75	0.13	0.74	0.66	0.56	0.84	0.54	0.48	0.48	0.83	0.90	0.81
VI	0.56	0.75	0.71	0.71	0.41	0.04	0.83	0.80	0.38	0.87	0.74	0.48	0.39	0.89	0.87	0.92
VII	0.65	0.77	0.68	0.74	0.16	0.29	0.55	0.46	0.72	0.62	0.91
VIII	0.38	0.51	0.53	0.50	0.02	0.53	0.51	0.37	0.80	0.49
Mean...	0.33	0.57	0.62	0.54	0.61	0.02	0.72	0.71	0.48	0.76	0.65	0.54	0.55	0.75	0.80	0.82
S.D.	0.22	0.11	0.14	0.18	0.12	0.09	0.08	0.07	0.17	0.11	0.07	0.12	0.13	0.17	0.09	0.15

identity of function probably accounts for the high correlation. However, it is a notable fact that several hours work in group intelligence tests yields a correlation with the reading ability composite which is not as high as 5 minutes work on the Burgess. The intelligence tests measure something more than, and something different from reading.

The Burgess test agree most closely with Directions ($r. = .76 \pm$ S.D. .11) and with Monroe Rate ($r. = 0.72 \pm$ S.D. 0.08) as might have been expected. It yields a correlation of 0.5 or better with any test save Brown's comprehension where the correlation is zero. The correlation with the Thorndike-McCall is lowest ($r. = 0.48$) with the exception just noted. This does not indicate an inadequacy of the Thorndike test, but rather indicates a fact, elsewhere verified, that the Thorndike test measures a quite different—even if correlated—function and that the two make an excellent team.

The correlations with the vocabulary tests are high; likewise that with Gray's oral. A wide reading vocabulary and mastery of the mechanics of reading is typical of the reader who excels.

(To be concluded in November.)

THE RESULTS OF RETESTS BY MEANS OF THE BINET SCALE

J. E. WALLACE WALLIN

Bureau of Special Education, Teachers College, Miami University

This study is based upon two testings in the St. Louis School clinic of 136 cases, three testings of 16 of these cases and a fourth testing of one of the cases. The retests were made at varying intervals. The average interval between the first and second tests for the entire number was 2.2 years, with an extreme range of from one-half year to six years. The average interval between the second and third tests was 2 years, based on the 16 cases given a third examination, while the range was from one year to almost four years.

The reasons prompting the requests for the re-examinations of these pupils were as follows: A few were re-referred to the clinic because they had been excluded from school owing to low mentality, and the guardians had applied for readmission. A few had been demoted to the kindergarten because of low mentality. When the eligibility requirement for admission to the special schools was lowered, they were referred to the clinic with a view to assignment to a special school.¹ Many such cases, however, were reassigned without further examination, because it was evident that a second examination was not needed. A few came from the special schools. The parents wanted them returned to the grades. But the large majority had been assigned to ungraded classes, and were referred because they failed to make adequate progress, either in the ungraded class in which they had been placed or in the regular grades in which they had been retained because an ungraded class was not available.² It is evident that this represents a highly selected group of subnormals³

¹ The authorities at one time fixed an intelligence age of six years as the entrance requirement for a feeble-minded child. This was subsequently lowered to five years. The present requirement fixed by state regulation, is a minimum intelligence age of about 3 years, or an IQ of about 30.

² The ungraded classes have been instituted for borderline, intellectually backward, and pedagogically restorable pupils, as explained in *Problems of Subnormality*, Chapter III, 1917.

³ The diagnoses made at the first examination were as follows: normal, 6; retarded, 12; backward, 48; borderline, 24; potential feeble-minded, 11; morons, 4; potential moron, 1; imbeciles, 12; idiot, 1; and deferred diagnosis, 17. All the categories except the last represent progressively graver degrees of intelligence deficiency. The average IQ at the first examination for those examined by the 1911 Binet was 0.79 (119 cases) and for those examined by the Stanford 0.61 (15 cases).

most of whom were reported for re-examination because they failed to make adequate progress in the ungraded or regular classes.

In view of the circumstances under which the children were examined, the expectation would be that the relative intelligence scores would show a progressive decline with each examination. What do the facts show?

Referring first to the diagnosis made, based on all the facts gathered on each case, we find that the second diagnosis was the same as the first in the case of 42 subjects, 14 of these having been diagnosed as backward, 10 as borderline and 10 as imbeciles. A higher classification was given to eight cases, and a lower classification to 70. The diagnosis was "deferred" the first time on 17 cases. We have not attempted to indicate here whether they were advanced or reduced in the later classifications. Of those given a lower classification, 7 were reduced one-half step,¹ 38 one step, 8 a step and a half, 4 two steps, eight two and a half steps, and one each three and a half and four and a half steps. Both of the latter who had tested normal or almost normal during the first examination, proved to be feeble-minded. The intervals between the examinations were 4.5 and 3.6 years, respectively. Twenty-six were reduced from backward to borderline (21) or to potential feeble-minded, 11 from borderline (9) or potential feeble-minded to feeble-minded, 8 from backward to feeble-minded, and 7 from retarded to backward. In the third examination three were given the same classification, and 11 were reduced, one one-half step, 6 one step, 3 two steps, and one two and a half steps, the diagnosis of the other two being deferred.

Turning to the more objective criteria, we find that the successive IQ's were higher for 12 subjects with the 1908 scale, for 16 with the 1911 scale, and for 7 with the Stanford scale.² The average amount of improvement between the tests for these subjects was 6.6 IQ in the 1908 scale, varying from 2 to 13 IQ; 4.3 IQ in the 1911 scale, varying from 1 to 8; and 8.1 IQ in the Stanford scale, varying from 2 to 17 IQ. It is evident that as determined by the measuring scales, some of these pupils improved their position. A few made unexpected advances, the number gaining seven or more IQ

¹ "Potential feeble-minded" and "potential moron" are counted as half-steps.

² It should be explained that many of the first and second examinations were made before the Stanford scale was in use and that, because of limitations of time, not all subjects who were tested after the Stanford scale was adopted were also given the 1908 and 1911 tests.

points amounting to 6 in the 1908 scale, 3 in the 1911, and 4 in the Stanford. The average improvement was greatest in the Stanford scale.

All of the others made lower IQ's in the successive tests, except 4 whose IQ's were the same in the 1908 scale, and 3 in the 1911. The average reduction in the second test amounted to 7.5 IQ in the 1908 scale (varying from 1 to 17); 7.8 IQ in the 1911 (varying from 1 to 23); and 5.1 IQ in the Stanford (varying from 1 to 10).

While the vast majority of these subjects suffered *relative* deterioration in intelligence (as measured by the IQ), practically all made *absolute* gains. The average improvement in the 1911 scale from the first to the second examination, for those (52) who were given the test both times, was a year and a half, the average time interval being 2.37 years, whence the annual improvement amounted to 0.42 of a Binet age. For those (8) who were given the 1911 a third time, the average improvement, during the interval of 2.25 years between the second and third examinations, was 1.1 year in terms of the scale, or 0.44 of a Binet age per calendar year. In the Stanford scale the improvement in intelligence from the first to the second examination, separated by a time interval of 1.87 years, was 0.67 year (based on 15 subjects who were retested by the Stanford), which is equivalent to a yearly improvement of 0.35 of a Stanford-Binet age, while the corresponding gain from the second to the third test, separated by an interval of 2 years, was 0.57 year (1 case). Not a single case showed a loss in intelligence age in any of the scales. In other words, not a single one of these children suffered actual dementia as determined by the scales, which may appear singular in view of the motley makeup of this group of cases, 2 of whom were epileptics, 2 choreics, 20 unstable and neurotic, 4 psychopathic, 29 speech defectives, 15 unruly and 21 wordblind (19 of these being dyslexia cases). These cases represent pretty much the "ragtags" of our run of cases.

So much for the extent of the gains and of the losses. We are also interested, however, in ascertaining how large the differences are between the IQ's of the successive tests irrespective of sign, *i.e.*, irrespective of whether the difference is a gain or a loss. Table I gives the average difference between the IQ's received in the first and second and in the second and third tests, and the average of these differences.

It will be observed that the average difference between the results in terms of IQ units, is not so very pronounced when the IQ's are

TABLE I
Average IQ difference between successive tests

Scale used	Between 1st and 2d tests		Between 2d and 3d tests		Between all tests	
	No. ¹	Ave. ²	No.	Ave.	No.	Ave.
1908.....	52	6.9	8	5.7	61 ³	6.6
1911.....	52	6.3	8	6.1	61	6.2
Stanford.....	15	6.4	3	5.6	19	6.1
1911 and Stanford.....	104	10.7	15	11.4	120	10.2
1908 and Stanford.....	104	14.	15	14.9	120	14.1

¹ Number of cases. ² Average difference between the IQ's irrespective of sign.

³ Includes the case given a fourth examination.

TABLE II
Successive IQ's for subjects showing the largest differences

No.	First examination				Second examination				Third examination			
	Chron. age	1908 IQ	1911 IQ	Stanf. IQ	Chron. age	1908 IQ	1911 IQ	Stanf. IQ	Chron. age	1908 IQ	1911 IQ	Stanf. IQ
1	8.16	98	96	..	9.16	92	90	..	10.66	87	87	
2	7.41	97	97	..	9.	93	93	..	11.75	87	84	
3	7.25	97	88	..	9.16	83	79	..				
4	7.25	1.00	94	..	8.58	93	86	..	10.58	..	84	
5	6.	93	89	..	8.08	1.07	1.07	..				
6	8.8	84	77	..	10.2	75	75	..	12.08	71	71	63
7	9.58	90	96	..	11.4	90	85	..	13.5	84	84	73
8	8.25	56	9.08	73
9	7.5	93	91	..	11.16	67				
10	8.91	88	80	..	11.58	71	69	66				
11	8.66	92	86	..	11.75	80	76	66				
12	6.75	74	74	..	9.75	87	82	66				
13	8.66	90	90	..	11.25	65				
14	8.25	76	76	..	9.66	50				
15	8.33	92	92	..	9.91	76				
16	8.16	1.07	98	..	12.41	75				
17	7.66	96	91	..	8.66	88	84	..	10.16	81	77	72
18	7.91	99	91	..	12.25	72	68	57				
19	8.5	33	10.1	45				
20	9.5 ¹	74	12.66	59				

¹ A recently examined case not included among the 136.

based on the use of the same scale. The difference in the successive testing amounts to less than seven IQ points. It is practically the same for the 1911 (Vineland) and Stanford revisions, and is not much higher for the 1908 revision. It is slightly larger for the first and the second tests, than for the second and third tests.

The difference is considerably greater when the first measurements were made by the 1911 or 1908 scales and the later measurements by the Stanford, for the reason that the Stanford grades lower, as will be seen presently.

Although the average differences between the successive IQ's are not very pronounced, whichever scale is used, individual instances occur in which the intelligence scores earned differ rather widely. Table II contains a score of such cases. Space permits reference to only a few of these.

CASE 6.—American, an only child. First examination, June, 1915, age 8.8:—Conduct good, but nervous, giggles constantly, chews ties and school materials, no concentration or retention, very poor in spelling and arithmetic, best in reading and singing. First steps at two, talked at about three. Father's family nervous and excitable, father and grandfather hard drinkers, father abusive to mother during pregnancy, and she "felt like committing suicide." Mother's people all "healthy." Physical examination: 4 dental caries and notched teeth. Recommended to an ungraded class and reexamination after a year, diagnosed as neurotic, deferred.

Second examination, November, 1916, age 10.2:—School report: cannot concentrate or retain, one-half year in kindergarten, 2 years in first grade, one half-year in second grade, repeating the work, greatest interest in animals, amiable disposition, takes correction kindly. Diagnosed as borderline, recommended to ungraded class, to which he had not been transferred, as no class was available.

Third examination, September, 1918, age 12.08:—School record: in III-3, doing I-4 successfully, best in spelling, poorest in number and reading, "good in physical and mental characteristics," "physically a fine upstanding boy," "beloved, good natured." Physical examination: looks normal, intelligent expression, four dental caries, conjunctivitis. Extremely deficient in reading, diagnosed as a moron, and assigned to a special school (had never been placed in an ungraded).

Between the first and the third examination he lost 13 IQ by the 1908, 5 by the 1911, 21 IQ by the Stanford as compared with the initial 1908 IQ, and 14 IQ as compared with the initial 1911 IQ. The report from the special school in June, 1920, indicates that his worst fault is lack of concentration, especially in academic work, he is very nervous and forgetful, not reliable, a tale-bearer, easily influenced, but generous and kind-hearted, he was doing second grade work in reading and spelling, and II-2 in arithmetic, greatest improvement in industrial work, in which, however, he shows little interest. Apparently the Stanford scale places this boy more accurately than the older scales.

CASE 8.—First examination, April, 1918, age 8.25. School record: sullen, lacking in concentration, easily discouraged, possibly doing kindergarten grade of work. Inmate of Masonic Home, family history unknown, except that father had pulmonary tuberculosis. Examination: stolid, vacant expression, right cervicals enlarged, fair physical condition, neurotic, lisps, lacramoseten dencies, diagnosis reserved, recommended to kindergarten.

Second examination, February, 1919, age 9.08:—School record: 1 year in kindergarten, one-half year in first grade, nervous, indolent, cannot concentrate, worst in reading, writing and all handwork, likes to work with dominoes. Examination: dull, infantile expression, one dental cavity, enlarged tonsils, stutters at times, lacramose tendencies still evident. Diagnosed as potential moron, and assigned to a special school. Report from school after one month: conduct good, but poorly developed socially, little self-control, infantile tendency to weep.

This boy made the significant gain of two years or 17 IQ by the Stanford in less than a year, and can be rated as not lower than a moron. To exclude such children from the benefits of the public schools on the basis of one Binet examination would be hazardous.

CASE 12.—Russian. First examination, October, 1916, age 6.75. School record: in I-1, but doing little, poorest in everything requiring thought, greatest interest in games, amiable. Examination: lisper, stupid reactions, diagnosed as potential feeble-minded, recommended for special school or ungraded class, and reexamination after a year or two.

Second examination, November, 1919, age 9.75. School report: in school 4 years, repeated kindergarten five times, I-1 twice, I-4 once and II-1 three times. In II-3, but spending two periods daily in an ungraded class, doing first grade work in language and arithmetic, poorest work in language, best in reading and arithmetic, greatest interest in games, baseball, and swimming, has tried hard to overcome speech defect and has improved, good natured. Spoke single words at one and a half years, phrases at two, but did not talk well until seven, according to the mother who said he "got better after his diseases" (measles at four and scarlet fever at nine). Examination: post-nasal obstruction (tonsils removed at three), vision $1\frac{5}{20}$ in each eye. Reads very well according to his intelligence level (Stanford 6.5), reading the Stanford ten-year selection in 30", with 3 misreadings and aid on "17 families," and reproducing nine memories. Diagnosed as potential feeble-minded, and recommended to a special school.

Here we find a curious disagreement between the old scales and the Stanford. When measured by the 1908 and 1911 scales there was an improvement of 13 IQ and 8 IQ, respectively, between the two testings. This improvement was transformed into a loss of 8 IQ in the Stanford scale. When comparing the Stanford rating with the 1911 secured on the same day, the Stanford shows a loss of a year and a half, or 16 IQ. We are satisfied that the Stanford rating is too low.

In the psychomotor test (Seguin) he graded about nine years, according to the writer's norms.¹

CASE 18.—Italian. First examination, June, 1915, age 7.91. School record: in ungraded class, best in reading, poorest in spelling and arithmetic, inattentive, unable to concentrate, varies from day to day, learns a little parrot fashion,

¹ Psychomotor Norms for Practical Diagnosis, 1916, Table XLIX.

physically strong, fair conduct, pleasant. Father stabbed a man. Physical examination, one carious tooth.

Second examination, November, 1919, age 12.25. School record: in ungraded, doing III-1 successfully, best in sewing, greatest interest in sewing, music, and industrial work, worst in arithmetic and language, erratic, quarrelsome, boisterous, stubborn. Physical examination: enlarged lymph glands, two dental caries, a trace of strabismus.

This girl tested normal by both of the old scales when first examined. Her subsequent history shows that this rating was entirely misleading. During a period of somewhat over four years her 1908 IQ fell 27 points, and her 1911 IQ 23, while the difference between the first 1911 IQ and the later Stanford IQ amounted to 34 points. We are persuaded that the Stanford scale rates this girl too low. She is by no means an imbecile, and barely, if at all, a moron. She read the Stanford Selection in only 29 seconds, with only two errors, but reproduced only $7\frac{1}{2}$ memories. In the psychomotor test she did as well as a ten and a half year old child. The instances in which the Stanford scale rates too low are quite numerous in our experience.

CASE 19.—American. First examination, May, 1919, age 8.5. School record: in kindergarten, greatest interest to run and play, unstable, troublesome at home. Aunt's brother went insane in army, subject to auditory hallucinations, strayed away. Spoke single words at four years; epileptic seizures since seven. Examination: inattentive, distractable, neurotic, slavers, occasional tendency noticed in left eye toward internal strabismus, adenoids and tonsils already removed, congenital lues suspected, intelligence age by Stanford 2.83, diagnosed as imbecile and excluded.

Second examination, November, 1920, age 10.1. School record: in school 4 weeks, in I-1, doing nothing, dribbles at times, health good, although struck by an auto a year ago injuring the head and spine. Physical examination: slight bilateral ptosis, slight stenosis in right nostril, restless neurotic, verbal perseveration, wandering attention, loquacious. Assigned to special school, where he was reported as doing kindergarten or subkindergarten work.

The interest in this boy is in the marked improvement which he made in a year and a half, amounting to 1.6 years or 12 IQ by the Stanford. With such a degree of improvement in an apparently hopeless low grade case—and we have school records of others evincing a similar gain—it is surprising that we have rejected the doctrine that all feeble-minded children should be denied the privileges of the public schools. The fact is that it is frequently impossible to determine for years whether a young mental subnormal is feeble-minded or not. We have examined a number of children who were unjustly excluded from school on the basis of a low test score and the assumption that the quotient would always remain the same. We counsel caution in the matter of the exclusion of assumed hopeless defectives from the public schools. The place in which to train

the mass of mental defectives is in the public schools, not the state institutions, for economic, if for no other reasons. The public schools exist for the impartial service of society's products.

In the following case the discrepancy in the intelligence rating was just as great, but in the opposite direction.

CASE 20.—American. First examination, February, 1918, age 9.5. School record: in II-1, doing I-3 successfully, greatest interest in drawing, raffia and writing, poorest in number, reading and spelling, good in conduct and disposition, very slow in comprehension, frequently absent because of bronchial cough. Examination: dental caries, myopia, neurotic, lisper, tendency to stutter, diagnosed as borderline and recommended to an ungraded class or a speech-correction class.

Second examination, March, 1921, age 12.66. School record: in ungraded, doing first grade in reading and spelling and third grade in arithmetic, after six years in school, best in arithmetic, sewing, and fitting together parts of wagons, toys, etc., greatest interest in carpentry, and mending school furniture, poorest in language, reading, spelling and writing, very helpful, always willing to assist, but has become rather rough with smaller boys since father's death. Examination: malformation of nose, speech much improved, decidedly deficient in reading, diagnosed as moron, and assigned to a special school. This boy only advanced five months in intelligence in about three years, while his IQ declined 15 points. The small gain in the Stanford is partly due, no doubt, to the literary character of the scale. The boy has little ability in language.

It is noteworthy that when examined on the same day by the 1908, 1911 and Stanford versions, all the subjects grade lower by the Stanford scale, except one who grades .2 year lower by the 1908, while two grade the same in the 1911 and Stanford scales. The average difference between the Stanford and 1908 scales amounts to 1.16 years and 11.1 IQ based on the first and second tests, the differences ranging from .3 to 2.1 years and from 2 to 20 IQ's (38 cases). Based on the second and third tests, the average difference amounts to 1.31 years and 12.7 IQ, ranging from .83 to 2.4 years, and from 7 to 29 IQ (8 cases). The average of these differences amounts to 1.19 years, or 11.4 IQ (46 cases).

The average differences between the Stanford and the 1911 scales for the subjects who were put through the two scales on the same day amount to .66 year (ranging from 0 to 1.5 years) and 7.3 IQ (ranging from 0 to 17—38 cases), based on the first and second tests; and to .90 year (ranging from .5 to 1.5) and 7.7. IQ (ranging from 1 to 13), based on the second and third tests (8 cases). The average of these differences is .71 year or 7.4 IQ (46 cases). The differences are greater between the second and third tests than between the first and second, possibly due to the increasing age of the subjects

and the absence of satisfactory tests in the higher ages in the 1908 and 1911 scale. The average chronological age of all the subjects at the first examination was 8.96 years, at the second 11.26 years, and at the third 11.69.¹ There was a marked increase in the average chronological age between the first and later tests.

It is apparent that the average difference in the age rating and IQ between the 1908 or 1911 and the Stanford is too large to be ignored,² while the differences in individual instances are occasionally surprisingly large, so large that quite contradictory conclusions would be reached according to the particular scale employed. We agree with the generally accepted conclusion that the 1911 scale is more accurate than the 1908 (Vineland), except in the highest ages, but the facts are not available whereby it can be conclusively affirmed that the Stanford norms are more accurate than the 1911,³ except in the upper ages, although the scale itself is much superior in various particulars. The necessity of validating the accuracy of the Stanford norms—and revising the tests and administrative procedure—on the basis of the testing of a large number of unselected children from various sections of the country is urgent. This should be done, in our judgment, by a disinterested committee of psychologists, several of whom should have had extensive experience in actual clinical work, the revised scale should be made available in an inexpensive edition at cost⁴ while the publication of reprints should be freely permitted without the fear of infringement of copyright. Our tentative conclusion, based on various considerations, is that most of the Stanford age norms are too difficult, thus exaggerating the subject's deficiency. Porteus has reached a similar conclusion so far as concerns the tests above age VIII.⁵

¹ The average age at the time of the second examination of the 16 examined three times was 9.69.

² The corresponding differences between the 1908 and 1911 scales were 4.5 IQ in the second examination (ranging from 0 to 12), and 5.3 IQ in the third examination (ranging from 0 to 19). The 1908 IQ's were lower than the 1911 in only 5 cases (during the second examination).

³ An incomplete analysis of the comparative accuracy of the two scales appears in *Preliminary Impressions of the Stanford Revision of the Binet-Simon Scale*, Psychological Clinic, 1918, 1 f.

⁴ Consider the vastly heightened cost of the Stanford revision and materials compared with the inexpensive Vineland guide, record forms, and materials.

⁵ Porteus, S. C. *Condensed Guide to the Binet Tests*, Training School Bulletin, 1920, 1 f.

MENTAL GROWTH AND THE IQ

LEWIS M. TERMAN

Stanford University

OTHER CONTRIBUTIONS ON THE VALIDITY OF THE IQ

Wallin¹ presents a criticism of the IQ based on data from Stanford-Binet tests of 411 backward and feeble-minded children in the public schools of St. Louis. His main criticism relates to the IQ distribution found in his various classificatory groups. The main groups were designated as "normal," "retarded," "backward," "borderline or potentially feeble-minded," "morons," "imbeciles," and "idiots." The author states that his classification of the subjects into these categories was based chiefly upon pedagogical and mental status, the latter determined by use of the Stanford Revision. Medical, social, and family data were also used. Nothing is stated with regard to how pedagogical status, intelligence test, medical and social data were weighted and combined. Presumably the final judgment was largely subjective, based upon empirical, offhand evaluation of the various kinds of data available.

The author informs us, however, that in no case was the IQ computed until after the diagnosis had been made.² After the classification was complete the IQ distribution in each group was examined. The extreme range of IQ for the different groups was as follows: "normal," 95 to 108; "retarded," 80 to 97; "backward," 59 to 94; "borderline and potentially feeble-minded," 56 to 84; "morons," 48 to 70; imbeciles," 21 to 65. That is, the range is wide and a large amount of overlapping is found. Hence the IQ is of no value for purposes of classification.

The argument overlooks two very important considerations.

In the first place, it is possible that about as much overlapping would obtain between Wallin's classification and that of another equally competent clinician using the same methods. The truth is

¹ J. E. Wallace Wallin: The Value of the Intelligence Quotient for Individual Diagnosis. *J. of Delinquency*, Vol. 4, 1919. pp. 109-124.

See pp. 146-147 of *The Intelligence of School Children* for my data on 183 re-tests of children above 110 IQ.

² One might infer from the author's discussion that life age beyond 15 or 16 was used as divisor in computing IQ's, although I can not be sure this was the case. As the subjects ranged in age as high as 19 years such a procedure would of course seriously affect the results.

that the names attached to these categories have as yet acquired very little exact meaning. There is little agreement either as to what they do mean or as to what they ought to mean. I should be surprised if the classifications by two clinicians, of the same children, were found to correlate more than 0.7.

In the second place, Wallin's classification is evidently not an *intelligence* classification at all. Just what it is, we are not informed, though in numerous articles he has made it clear that he considers various other factors as important as intelligence in the diagnosis of feeble-mindedness. On p. 124 of the article in question he defines feeble-mindedness purely in terms of social and vocational incompetency. This is a common use of the term and is of course legitimate for practical purposes. However, as I have elsewhere pointed out,¹ it is a concept of little use to science. One's ability to get on in the world depends upon an indefinite number of accidental factors, including health, looks, inherited wealth, friends, local industries, the economic condition of the country, etc. Surely no one ever supposed that feeble-mindedness, in the sense of social incompetency, is accurately measured by the IQ.

Finally, the author takes too literally the IQ classifications others have proposed. As for my own classification of children as normal (IQ 90-109), dull (80-89), borderline (70-79), feeble-minded (below 70), etc., it never occurred to me that any one would construe this as marking off well-differentiated groups, or as intended for anything more than a rough tentative classification. The known probable error of an IQ score would itself make any such rigid classification quite absurd. I myself have pointed out² that even if we had a perfect measure of intelligence we could not expect it to furnish an absolute index of an individual's educational or social success. The following statements by me (p. 87, *The Measurement of Intelligence*) is explicit on this point:

"It must be emphasized, however, that this doubtful group is not marked off by definite IQ limits. Some children with IQ as high as 75 or even 80 will have to be classified as feeble-minded; some as low as 70 IQ may be so well endowed in other mental traits that they may manage as adults to get along fairly well in a simple environment."

¹ Lewis M. Terman: *The Binet Scale and the Diagnosis of Feeble-mindedness. J. of Criminal Law and Criminology*, Vol. 7, 1916. Pp. 530-543.

² *The Intelligence of School Children*. Pp. 97-110, p. 127 ff; also *The Measurement of Intelligence*, p. 80-81.

Elsewhere¹ I have stated my conclusions on this and related matters still more fully, in particular pointing out: (1) that no intelligence scale gives an entirely accurate measure even of intelligence; (2) that in the diagnosis of feeble-mindedness, medical, neurological, and social data are necessary; (3) that social competency and educational possibilities both depend largely upon non-intellectual mental traits; and (4) that no responsible psychologist would think of using a Binet test score as an automatic criterion of feeble-mindedness. "If any psychologist ever hoped to find such a simple standard as 12-year intelligence (or 75 IQ, etc.) an infallible criterion of fitness to be at large, surely he has long since been disillusioned. The writer does not for a moment suppose that those who have proposed such standards ever meant that they should be rigidly and mechanically applied" (p. 536). Continuing, I pointed out that the term feeble-mindedness is currently used in two very different senses, one referring to intellectual defect, the other to social or vocational incompetency. "Intellectual feebleness," being a fairly definite thing and at least roughly measurable, is a term usable in science; "feeble-mindedness" in the sense of "social in efficiency" is not.

The limitations of the IQ have also been made the subject of a spirited article by Dr. Mateer.² Data on fifteen specially selected cases are presented to show that the IQ does not always remain constant and that it can not safely be taken as a basis for differential diagnosis. "The IQ or C. I. A. of the individual feeble-minded child is sometimes not in the least a factor differentiating him from normal children of his age. His IQ may decrease steadily through even the earlier years of childhood, it may stand still, it may even temporarily increase. Even an IQ of 75 or 70 or 60 need not mean feeble-mindedness. It may mean dementia, either in the sense of insanity or of other deteriorating neural condition, as for instance a juvenile paresis."

I do not think anyone would dispute Dr. Mateer's contention that the IQ does not *always* remain constant, especially in the case of psychopathic subjects. Even as regards normal subjects its constancy is never, so far as I know, referred to as anything more than "relative" constancy, "approximate" constancy, "a tendency to" constancy, etc. Everyone makes liberal reservations regarding its

¹ *J. of Criminal Law and Criminology*, 1916. Pp. 530-543.

² Florence Mateer: *The Diagnostic Fallibility of Intelligence Ratios*. *Ped. Sem.*, 1918. Pp. 369-392.

constancy as far as epileptic, insane, or other types of psychopathic subjects are concerned.

Dr. Mateer gives no data for normal children, and her conclusion, "it is self-evident that IQ's do not remain constant" (p. 385) can not be taken as demonstrated to be the rule even for feeble-minded children. As a matter of fact, the fifteen cases which she reports are nearly all admitted to be psychopathic. For example, the clinical descriptions of the cases are full of such phrases as the following:

Case 1. "Psychotic." "May develop an actual insanity."

Case 2. Has "always been peculiar." "Convulsions," "head-aches and pains." "Possibly epileptic."

Case 3. (Tests normal.) "Neuropathic family." "Steals," "lies," is "obscene." "He is, undoubtedly, of a neuropathic predisposition."

Case 4. History of "killing animals." "Likes blood." "Already gives evidence of neural disturbance." (Later test showed degenerative changes.)

Case 5. Handicapped by "neuropathic instability." Tests nearly normal and is bright in school, but "masturbates," "lies," and is "listless," "unreliable," etc. (The author suggests that the abnormal mental disposition may be connected with an enuresis which was present, possibly due to failure to form right conditioned reflexes in childhood.)

The other ten cases, five of whom tested below 76 IQ and even below 80 IQ, are said to be "undoubtedly feeble-minded," though one is reported as "bright in school." "Four are cruel to children, two are cruel to animals to the extent of putting cats on hot stoves. At least three of them are pyromaniacs, eight have temper spells, seven of them are destructive. In several cases the parentage is unknown, but the rest are weighed down by a history of alcoholism, a little insanity, general inferiority *but not feeble-mindedness*, and three of them are illegitimate." (*Italics mine.*)

Dr. Mateer seems to hold it against the intelligence tests that all these children once tested at age. The initial tests, however, were made by the Goddard Revision. I calculate that by the Stanford Revision none of the initial IQ's would have been above 90. In fact, the child who gave the highest initial IQ found by the Goddard Revision (103) was also given the Stanford Revision, with a resulting IQ of 90. If we make reasonable allowance for the scale used there are only five of her fifteen cases which show more than about 8 points of change in IQ over a period of one to five years. Probably anyone with

considerable clinical experience could easily duplicate this handful of exceptional cases described by Dr. Mateer. It may be done any number of times without destroying the usefulness of the IQ for such purposes as it is really intended to serve.

Of course it is unfortunate that the IQ does not enable us to diagnose psychopathy, epilepsy, enuresis, etc., or tell us whether the subject has or has not formed the appropriate conditioned reflexes. For such purposes, one is bound to admit, the IQ is distinctly fallible. It might be well to warn astronomers that there may be similar limitations to its usefulness in the prediction of eclipses!

In a more recent article on the "Interpretation and Application of the IQ"¹ Professor Freeman has approached the problem from a different angle and has raised some important questions. He rightly observes that the validity of the IQ hinges upon the greater overlapping of mental ages in the upper years than in the lower; that it would require, for example, the standard deviation of mental ages of unselected 10-year olds to be about twice that for 5-year olds, and the standard deviation for 15-year olds to be about three times that for 5-year olds. He notes that those who have used Binet tests with unselected children have usually found such increase in mental age overlapping. When he examined the results of group intelligence tests, however, no such rule was found to obtain. Data from several group tests are presented, and in every case the variability, expressed in terms of point score, shows a tendency to remain constant between the ages of 7 or 8 and 12 or 13.

I think there is no question about the correctness of Professor Freeman's observation. I found the same thing three years ago for army test Alpha, and have since found it to hold also for the Otis, National, and Terman group tests. For example, in the case of unselected children of 8 to 14 years (about 175 at each age), the standard deviation of total score of Scale A and Scale B of the National Test remained almost constant at about 50 points through this entire age range.

We thus have an apparent contradiction, and unless it can be shown that one or the other of these findings is not in accord with the facts, some explanation must be sought which will harmonize them. I believe that further investigation will confirm the essential correctness of both findings. As far as the Binet results are concerned, the

¹ *J. of Educational Psychology*, Vol. 12, 1921. Pp. 3-13.

progressive increase in mental age overlapping is shown in the data of various investigators. For example, in my tests of 905 unselected children the interquartile mental age range of 6-year olds was 10 months, and of 12-year olds 20 months. Similarly, Bobertag¹ found 11-year olds to overlap 12-year olds on Binet tests almost twice as much as 6-year olds overlap 7-year olds. Striking confirmation of these results is found in Burt's report on "The Distribution and Relations of Educational Abilities" in the case of a representative group of 31,965 London school children.² In this study it was found that "—in educational ability normal children tend to vary [using the standard deviation as the unit] above and below the average level for their age as follows: at the age of 10, by at least 1 year; at the age of 5, by just 0.5 of a year; at the age of 15, in all probability, by nearly 1.5 years, and throughout, by about one-tenth of their age." (p. 31.) The standard deviations for the ages at which the subjects were considered representative were as follows:

Age.....	4.0	5.0	6.0	7.0	8.0	9.0	10.0	11.0	12.0	13.0
Stand. Dev.										
(years).....	0.34	0.55	0.62	0.63	0.75	0.91	1.10	1.17	1.18	1.24

The apparent contradiction might be explained as due to the shifting of the point score values of group tests in the different ranges of the scale. Perhaps no one would claim that a point in score has equal value over the entire scale range. On the other hand, one would hardly expect the shift in score values to be large enough to account for the phenomenon in question. Another explanation is suggested by a consideration of the psychological differences between the Binet scale and current group tests. The latter endeavor to measure the same functions throughout the range over which they are applied. Binet tests, on the other hand, to a great extent measure different functions at different levels. It is constructed on the theory that mental growth does not imply equal development of all the particular capacities at once, or in the same particular capacities at all periods; that certain differences in mental functions appear in a more or less definite order. It is adapted to bring out the fact that the 14-year old, for example,

¹ Otto Bobertag: Die Intelligenzprüfungsmethode von Binet-Simon bei Schwachsinnigen Kindern. *Zeitschrift f. ange. Psychol.*, 1912, 6. Pp. 495-538.

² Report by the Education Office submitting Three Preliminary Memoranda by Mr. Cyril Burt, M.A., Psychologist, on the Distribution and Relations of Educational Abilities. London, P. S. King and Son, 1916. Pp. 93.

excels the 7-year old not merely in the maturity of certain mental functions, but that he is mentally able to do various kinds of things which 7-year olds can not do at all. The group tests, being so much more restricted, probably fail to bring out these differences to so great an extent as does the Binet, and as a result give a variability in the upper ranges which is less than it ought to be. In this respect they do not afford an entirely satisfactory basis for the psychological analysis of mental growth changes.

My suggestions, however, are only tentative, and I shall hope to return to the question at another time. Professor Freeman has raised an issue of real importance.

CRITERIA TO EMPLOY IN CHOICE OF TESTS

RAYMOND FRANZEN

Director of Research of the Public Schools of Des Moines, Iowa

AND

F. B. KNIGHT

Asst. Prof. Educational Psychology, University of Iowa

Instruments employed in the exact determination of the quantities of abilities and capacities involved in school procedure, have multiplied in the last years until the predicament of an administrator is no longer to find a test but to choose which of existent tests he will use. When we are in need of indices of reading ability, we must decide the comparative value of the many tests which purport to measure reading in order that we may express our diagnoses in the best medium available. Never a day passes that some officer of public instruction does not decide to use some one test for some one purpose. What criteria have influenced his choice?

Geographical preferences are not economical; the mails will allow transmission of tests from one area to another. And still one reading test is predominant in the west and another in the east. Advertisement should bear no weight in the dissemination of tests; we ought to be sufficiently familiar with all the available material to decide the value of tests independent of their commercial publicity. Nevertheless some group intelligence tests are being used in preference to others less advertised where the data at hand does not substantiate the choice. A test is not justified solely by the perspicacity and ingenuity of its maker,—the original data and the technique of construction are in most cases available,—and yet the prestige and influence of the author are often the sole bases upon which decisions of the comparative values of tests are assigned. This triumvirate of criteria—geography, advertisement and prestige of author—we should discharge from our educational judiciary.

Another triumvirate—administrative exigencies—needs to be given a less emphatic voice than it now exercises. The price of a test, the time it takes to give it and the convenience of scoring facilities are important factors in the choice; but only if all the tests considered perform the service which is needed. They are secondary criteria. As soon as we know how much service we can expect from each of a group of tests, and not before, can we decide whether or not we can afford the time and money necessary to their administration. If a

test does not really measure reading, and if it is reading evaluation which is the objective, then it would pay an administrator to use the test that did this service even though it cost more, took longer to give and was scored at the expense of a great deal of time and energy. Often he would be better off if he considered first the other criteria pertaining to the psychological and statistical values of the test, *because he would then give no test at all*; that is, often the only test he can afford to give is the only test which does not perform the service he is seeking. Let him decide primary values first; then let him consider price, time and convenience in ratio to the benefits derived.

Discarding these inexact, irrelevant and secondary considerations, what criteria shall be employed? A perfect test may be used in many ways and therefore has virtues which are for some of its uses superfluous. We will enumerate all of these virtues and then outline the relative value of each for the most important uses of measurement in school life. A test may to a varying degree:

1. Measure what it purports to measure.
 2. Yield the same diagnoses tomorrow that it does today,—the reliability of a test.
 3. Yield the same diagnoses in the hands of one examiner that it does in the hands of another,—the objectivity of a test.
 - 4a. Yield numerical diagnoses, the units of which are equal, so that equal numerical increments are symbols of equal increments of the ability measured,—the scaling of a test.
 - 4b. Mean nothing at all of the quality measured by the zero of its scale, so that a score of eight is twice a score of four etc.
 5. Provide standards by which comparisons may be made to large numbers of any one grade and of any one age,—the norms of a test.
 6. Interest the child.
 7. Register a wide range of abilities.
 8. Distinguish between failures, so that we can tell *why* a child has a low score as well as *that* he has a low score.
 9. Correlate to unity with intelligence when the abilities measured are at their maximum. (It is obvious that for tests of some abilities this is not desirable, for instance, mechanical ability.)
- NOTE.—No credit is here given or taken for originality in the formulation of criteria. It is hoped that this is a convenient assembly of important test virtues.

The needs which prompt an administrator or director of research to the use of tests can be classified into five main heads: (1) Compari-

son,—with other cities, schools within the district, or individuals; comparison of either the central tendencies or the spread; (2) Experimentation,—to find the value of curricula, methods, text books or time allotment; (3) Classification,—by intelligence and by information; (4) Diagnosis and prognosis,—including the comparison of degrees of attainment with measurements of potentiality; (5) Definitive outline of goals,—qualitative definition in terms of tests, quantitative definitions in terms of locations on the scales of those tests.

A test must measure what it purports to measure to be employed profitably in any of these ways. Our first criterion applies then to all the uses. The test must either do this on the face of it or have a known correlation with a known criterion to vouch for its authenticity. The reliability and objectivity of a test are important considerations too in each of the uses of tests. The reliability of an average and of a measure of spread are functions of the reliability of the test. Comparisons of the average or variability of a group in two abilities is not permissible unless we know the reliability of the tests; the use for experimentation of a test with a low reliability leads to faulty conclusions and classification; diagnoses and definition of goals made on a basis of unreliable material is tantamount to prescriptions of a doctor who has made a diagnosis over a telephone. Remedial treatment implies reliable diagnoses.

Whereas it is always of great value to be able to compare scores, knowing that a unit on any portion of the scale is equal to a unit on any other portion, to be able to say that a score of 87 is just as far above a score of 82 as a score of 30 is above 25, it is a *sine qua non* in the use of a test for most experimental purposes, since progress along a scale is generally involved. If we cannot compare progress, and we cannot unless a test is scaled, then we cannot gain much from comparison, experiment, classification, diagnosis or quantitative definition. Scaled tests are always better than tests whose units are undefined; in many situations no test at all does less harm than the use of an unscaled test, scores on which are interpreted *as though they were scaled*. Teachers and administrators generally do interpret scores so.

Standards are important in order that we may compare. We are sufficiently awake to this need. It needs emphasis that we could use standards of variability as well as standards of central tendency. We should be able to compare the spread of the abilities of our 5th grade with normal spread as well as the comparison of the average equipment in their possession with normal equipment. We know no test on the market today with published norms of spread. It would be very

valuable for us to know standard deviations of the tests we use. The average reading of 5th grades in Des Moines may equal average 5th grade reading in New York, and yet our standard deviation may be twice as large. As this would indicate a wide disparity of attainment among our 5th grade children, we would want more to know of it, coupled with a comparison of our S.D. in intelligence to normal S.D., than we would want to know the averages here and elsewhere. For convenient diagnosis and classification, age norms are necessary. Then we can translate scores in any measured ability into indices of maturity and make use of age as a common denominator to gain inter-comparison of all abilities and capacities. For experimental purposes standards are, of course, often unnecessary.

That a test is better for purposes of comparison, classification, diagnosis, prognosis and quantitative definition if it interests the children and if it is applicable to widely separated extremes in degree of possession of the quality under investigation, is readily comprehensible. The more the children are interested the easier it is to obtain optimum results. The wider the range which the test measures the more you can compare, the further you can predict and the more inclusive is the definitive outline. It is obvious that if we can use the same test from 3rd grade to senior high school, that will be better than using one from 3rd through 5th, another from 6th through 8th and still another from 9th through 12th. For experimental purposes these two things may not be desirable, and in some cases they may be undesirable;—it is conceivable that a test may be chosen because of its lack of interest to suit certain experimental conditions.

A test is always better than another, other things being equal, if it distinguishes between failures. It is good to be able to say which children have a low "Arithmetic ability." It is better to say that a child has a low adding ability in the fundamentals. It is best to say a child has a low adding ability in the fundamentals because his combinations of 9 and 5 are weak. If we can tell why a person is weak in terms of the elements that contribute to strength, then remedial work is readily encompassed. Much work on our tests needs still to be done before we reach an ideal in terms of this criterion.

The last of the listed criteria also applies to all but the experimental use. Its value is that it affords the possibility of comparison of achievement in a function to the intelligence involved. It provides a direct check on the correspondence of the two axes of classification—capacity and information, a diagnosis in terms of inherited capacity and a definition of goals in terms of the intelligence of the children for

whom the goals are instituted. In a word, if we know the most that can be, we can better deal with the amounts that we now have. Nature has made lavish investments in the nervous systems of a few. These investments should be made to pay social dividends. We can only accomplish this through ratios of achievement to intelligence. These ratios demand high correlations of product tests with intelligence tests when the abilities are pushed to their limits.¹ In case the test being judged is an intelligence test the criterion still holds.

In closing it may be appropriate to emphasize the fact that the first criterion is not a pedantic insistence on the obvious. There are many tests on the market which seem to test an ability and which in reality do not. An "arithmetic" test, for instance, which lays undue stress on time might test speed in writing figures, a "reading" test which had in it too many artificial difficulties not ordinarily encountered in reading might test attention. These facts have been well pointed out by Thorndike and Courtis² and again by Pressey and Pressey.³ A choice of a test to perform a function should be preceded by careful study of all criteria and where published data, correlation with a criterion, reliability coefficients, correlations with other tests, distributions of unselected data and scale determinations are not available one should think twice before using the test.

We should insist upon the use of tests which can be proven to test what they purport to measure, which are reliable and objective, which are scaled and which have well defined norms based on sufficient material. If in addition we are able to select tests which interest the children, are applicable to all grades, distinguish between failures and correlate highly with intelligence at their maximum, we will be able to manipulate our results in such manner as to gain additional benefits. Certainly a consideration of these test virtues will avoid much useless and even harmful work in survey of abilities and may contribute toward a selection of the better instruments of measurement by diminishing the sales of tests which readily confess their inadequacy to a judgment in terms of these criteria.

¹ Refer to Raymond Franzen: *An Accomplishment Quotient*, Nov., 1920. T. C. Record.

Rudolf Pintner and Helen Marshall: *A Combined Mental-educational Survey*, Jan., 1921, *Journal of Educational Psychology*.

² Thorndike, E. L. and Courtis, S. A.: *Correction Formulae for Addition Tests*, T. C. Record, 1920, 21, 1-24.

³ Pressey, L. W. and S. L.: "A Critical Study of the Concept of Silent Reading Ability," *Journal of Educational Psychology*, Jan., 1921.

PERSONAL JUDGMENTS

E. E. LINDSAY

The State University of Iowa

The purpose of the study herein reported was to compare teachers' estimates of childrens' native capacities with these capacities as determined by as scientific a method as possible. The scientific method used was the Binet-Simon tests; the teachers were a group of graduate students together with two University professors; the children were a tenth grade history class. The class was made up of twelve girls and seven boys, coming from widely varying home conditions. The graduate students were all men of special training in education and, with one exception had had years of teaching experience. The Binet-Simon test is assumed to be a fair measure of the individual capacity.

After a class-room acquaintanceship of at least one month, the seven members of the teaching group were asked to rank the class in the order of their native mental capacity. In this judgment they were asked to eliminate all such factors as personality, effort, attainment, etc. The results of these independent judgments together with the examination grades and the IQ's are ranked and presented in Table I:

TABLE I

Pupil	IQ	Exami- nation score	Ranks									
			IQ	Exami- nation	Regular teacher	Pro- fessor	V	W	X	Y	Z	VWXYZ
1g	122	85.0	1	4.0	5	1	2	2	3	2	2	2.5
2g	109	83.0	2	5.0	4	11	17	15	7	12	13	12.0
3g	102	80.0	3	6.0	10	9	3	3	4	7	5	4.0
4g	100	92.0	4	2.0	2	2	1	1	2	3	3	1.0
5g	99	77.0	5	7.0	13	6	10	5	8	5	4	5.0
6b	95	65.0	6	8.5	3	7	7	9	16	4	8	8.0
7b	95	97.0	7	1.0	1	4	4	4	1	1	1	2.5
8g	93	90.0	8	3.0	6	5	6	10	6	6	6	7.0
9b	93	65.0	9	8.5	8	17	13	18	12	15	19	18.0
10g	93	50.0	10	14.5	12	18	11	8	9	9	9	9.0
11b	90	46.0	11	16.0	18	12	18	13	13	17	15	16.0
12g	88	63.0	12	11.5	7	3	5	6	5	10	7	6.0
13b	88	62.0	13	13.0	15	8	12	7	10	11	10	10.0
14g	87	39.0	14	17.0	17	14	8	16	11	16	11	11.0
15b	81	64.0	15	10.0	9	10	9	11	17	19	12	13.0
16b	77	50.0	16	14.5	11	15	16	12	15	13	17	14.0
14g	75	35.7	17	18.0	19	16	19	17	14	8	18	16.0
18g	68	29.0	18	19.0	14	13	15	14	19	14	14	16.0
19g	66	63.0	19	11.5	16	19	14	19	18	18	16	19.0

This table is read—Ig's IQ is 122; her examination score is 85 per cent. Of the 19 children she ranked: in IQ—1st; in examination score—4th; in teacher's judgment—5th, . . . in composite judgment of the five graduate men—2.5. The ranks and scores underlined fall within the range of normal on the IQ basis, namely 90 to 110. Sex of pupils is indicated by letters *g* and *b*.

Table I reveals wide discrepancies between native capacity as measured by the Binet-Simon test and by personal judgment. 2g's case perhaps is the most extreme. Her IQ ranks second in the group and yet she is placed as far down as 17th, or sub-normal, by one of the men, while by no one except her regular teacher is she placed higher than seventh. The group as a whole placed her more than half the total range below her IQ position. 9b is another extreme example. In opposition to these differences there are cases like 5g's where the group and the IQ rankings are very similar.

As a better method of comparison the correlations between each of these rankings with all of the others were compiled using the Spearman R method.¹ This table follows.

This table reads, IQ's correlate with examination grades 0.53; with judgments of individuals as follows: regular teacher—0.38, professor—0.43 . . . , composite judgment—0.52.

TABLE II

	Exami- nation	Regular teacher	Pro- fessor	V	W	X	Y	Z	VWXYZ
IQ.....	0.53	0.38	0.43	0.40	0.47	0.52	0.45	0.47	0.52
Examination...	0.61	0.43	0.46	0.36	0.48	0.42	0.47	0.46
Reg. Teacher...	0.38	0.48	0.33	0.35	0.43	0.42	0.38
Professor.....	0.50	0.63	0.47	0.48	0.60	0.64
V.....	0.55	0.48	0.42	0.63	0.62
W.....	0.48	0.53	0.68	0.76
X.....	0.52	0.58	0.66
Y.....	0.62	0.63
Z.....	0.82

With one or two unimportant exceptions the correlations displayed by this table are none of them high enough to be significant to any

¹ This work was done by a class in statistics under the direction of Dr. H. A. Greene of the College of Education, State University of Iowa.

marked degree. They would indicate the examination grade as the best criterion of native ability but even here the correlation is not high enough to warrant the assumption that scholastic standing, as determined by examinations, is a safe guide to mental endowment. It is interesting to note that of the three individuals whose judgments correlated lowest with the IQ, two were university professors and the other the youngest graduate student with no teaching experience. The regular teacher's estimate correlated the lowest of all. The high and low judgments correlated with each other 0.35.

In drawing conclusions from this experiment, two factors must be considered. The group judging was a very highly selected one, both as to training and experience, and the number of cases involved is small. The findings would tend toward the following conclusions.

1. Teachers' estimates of children's native capacity are significant, but to no marked degree.

2. Training and experience of the teacher do not seem greatly to affect this significance.

3. Individual judgment of the same children by observers with approximately the same contact differ widely.

4. Other factors than native ability *do* enter into one's judgment of same.

NOTES ON ARTICLES IN EDUCATIONAL PSYCHOLOGY IN CURRENT ISSUES OF OTHER MAGAZINES

REPORTED BY CECILE COLLOTON

Department of Educational Psychology, Lincoln School of Teachers' College

EDUCATIONAL TESTS

Cooperative Chemistry Tests. Seth Hayes. Journal Educational Research, 1921, Sept., 109-120. Description of the work of chemistry teachers in standardizing questions for the subject of chemistry as presented in Cleveland.

The Measurement of College Work. Ben D. Wood. Educational Administration and Supervision, 1921, Sept., 301-331. Report of an experiment conducted by the staff of instructors in Contemporary Civilization in Columbia College: a new method of examination and its results.

Interpreting Achievement in School in Terms of Intelligence. I. N. Madsen. School and Society, 1921, July, 59-60. A method of computing an Achievement Quotient based on age-grade standards in intelligence and educational tests, to show the relation of actual achievement to possible achievement.

INTELLIGENCE TESTS

On the New Plan of Admitting Students at Columbia University. Dean H. E. Hawkes—Dr. A. L. Jones. Journal of Educational Research, 1921, Sept., 95-101. The mental test as a possible substitute for the old method of entrance examinations.

Intelligence Classification and Mental Hygiene. Garry C. Meyers. Pedagogical Seminary, 1921, June, 156-160. A scheme for a nationwide classification of school children on basis of intelligence rating; its practical advantages in terms of higher mental and social efficiency.



A Ten-minute Intelligence Test in Junior Employment Offices. Harold H. Bixler. School and Society, 1921, Sept., 166-168. Comparison of ten-minute Test Z (used to test all applicants at the Pittsburgh Public School Employment Office) with the 45-minute Otis Test. Coefficient of Correlation = 0.7.

The Reliability of the Binet Scale and of Pedagogical Scales. Arthur S. Otis and Herbert E. Knollin. Statistical Formulæ for determining the reliability of scales.

LEARNING IN THE SCHOOL SUBJECTS

A Year's Study of the Daily Learning of Six Children. George E. Freeland. Pedagogical Seminary, 1921, June, 97-115. Factors in learning as shown by an extended study of six normal children, grades 1 to 6, learning to typewrite under normal school conditions.

NEW PUBLICATIONS IN EDUCATIONAL PSYCHOLOGY AND RELATED FIELDS OF EDUCATION



A New Book for Teachers of History.—This book, says the author in his preface, has been written in the interest¹ of better history teaching. While the technic of teaching has received chief emphasis, the book contains many excellent suggestions and concrete illustrations covering the problem of organizing history courses for teaching purposes.

From the point of view of educational method, the book discusses the history recitation. In this connection it treats of the textbook, source, topical and problem methods of presenting subject matter. It gives suggestions for teaching pupils to study history. It illustrates the problem of written work in history as well as taking up in detail the question of the collateral reading of the pupil and his use of the high school library. It gives some space to the new movement to standardize tests and examinations in history; though this treatment would have been more helpful if it had told in detail how the teacher was to use these tests. Illustration of their value in diagnosing individual and class difficulties might well have been included. The book also takes up the question of teaching current events in connection with history.

Chapter eleven of the book which discusses the planning of the course and the organization of daily lessons and should prove to be particularly helpful to the beginning teacher. Here the author outlines a scheme of general organization into distinct periods of history. Within each period he illustrates types of the daily work, *i.e.*, he gives examples of outlines, of maps to make, collateral reading, dates and events to know and to remember, and historical personages to know and to identify. These important details covering the organization of the course might have been treated more at length for it is here that the inexperienced teacher needs much assistance.

It is to be regretted that the author did not discuss more in detail psychological and pedagogical phases of the subject. For example

¹ Tryon, R. M. *The Teaching of History in Junior and Senior High Schools*. Ginn and Co., 1921. 284 pp.

the psychological processes involved in the learning of history the place of association, the amount of repetition necessary to fix connections and the use of imagination and judgment are important questions of educational method with which any history teacher should be familiar. Such mooted questions as the proper treatment of important institutions, description of the life of the times and the relating of history to important contemporary activities and problems should also be treated in some detail in a book on the teaching of history.

The reviewer agrees with the statement in the preface that while there are many ways of teaching history it is fundamental to educational method that the teacher know that there are a number of ways of doing the multitude of things connected with everyday procedure in that subject. This book tells the teacher of history about many of these ways and points out the procedure in managing them.

EARLE RUGG.

The Army Mental Tests.—Much has already been written of the work of psychologists in the United States army during the recent war. Their services were manifold and psychologists were utilized in many different branches of the army. The present volume¹ deals with the work of the group of psychologists who were engaged in giving mental tests. This was done under the direction of the Surgeon General's Office. The work started shortly after the United States declared war and continued long after the armistice. Hundreds of psychologists were engaged in the service and nearly two million men were examined. The reader will, therefore, readily appreciate the difficulty of telling the story of this vast undertaking and of presenting the results of such a large accumulation of tests. The volume is under the general editorship of Yerkes. It is divided into three parts, each under a separate editor, namely Yerkes, Terman and Boring respectively, and each one of these expresses obligations to many helpers.

Part I deals with the history and organization of the service from the first unofficial tests tried out under the auspices of the American Psychological Association, through the official trial of the plan, the organization of the service as a part of the military establishment, down to the abandonment of the work in 1919. The report brings

¹ Psychological Examining in the United States Army. Edited by R. M. Yerkes. Memoirs of the National Academy of Sciences. Vol. XV, 1921. Pp. 890.

out well the various difficulties that the psychologists had to meet and it makes a very instructive, although at times, depressing picture. What seems to have retarded and hampered the work more than anything else was the continual confusion of psychological examining with psychiatric work. It emphasizes again, what was apparent at the time, that the psychological service was misplaced as a part of the medical branch of the army, that it should have been an integral part of personnel work, and not under the control of the Surgeon General's Office. In spite of these and many other handicaps, it is gratifying to know how much was accomplished and how favorably in general the work was received. The whole report shows clearly the perseverance and doggedness on the part of the Chief of the Division of Psychology in the face of much prejudice and ignorance. He and his co-workers are to be congratulated upon what they accomplished.

In Parts II and III we are given a description of the development of the tests employed and some of the more important results. Part II is in the main historical and deals largely with examination *a*, from which *alpha* was developed. It also explains how the need for a test of illiterates and foreigners arose and how in response to this *beta* and the performance scale were constructed. Part III presents distributions of scores for a fair sampling of the two million men tested. There was neither time nor opportunity to tabulate the complete data, nor is it likely that much additional information would have been obtained by so doing. There is in this part of the book a great number of distribution tables and it will prove a perfect mine for the statistician who may desire to work up other aspects of the data. The conclusions drawn are extremely conservative and they do not go beyond the data at hand. They arouse in the mind of the reader many interesting conjectures, and these results should prove a stimulus for further research in many directions.

Because of the very nature of a large work such as the one before us, it is inevitable that there should be a certain amount of repetition and overlapping, and this is the case particularly with reference to Parts I and II. There are several lines of investigation that one would have liked to see undertaken or discussed more thoroughly, but it would be absurd on the part of a reviewer to point out omissions, because the editors are keenly conscious of such themselves and could unquestionably tell of more things that might have been done than any reviewer could. Considering the time and assistance available, they have certainly made the most of their opportunities.

The reader of this volume is impressed with the amount of work accomplished in so short a time. It is a splendid example of what cooperative research can do with the right motive or stimulus. The army testing established the validity of the group test method in the space of a few months and silenced the doubts and misbelief that were current before that time. It furthermore has given us the best estimate of the average mental ability of the population at large and has shown us how much we had overestimated this in the past. And, lastly, it brought the use of mental tests very much to the front and demonstrated their value to the world at large. It is very valuable to have a complete record of the work. This volume will remain a worthy monument, better than one in stone or brass, recounting the patriotism of American psychologists and their desire to serve their country in times of great need.

R. PINTNER.

Adolescence.—Ever since Stanley Hall wrote his monumental work on adolescence, this period of life has attracted many writers. The present book¹ is by the author of the well-known "Psychology of Childhood," which was published in 1893 and which formed an important contribution at that time to the child-study movement. In the same way the author has now given us the benefit of his observations and thoughts on adolescence. In a pleasant, readable style he discusses instinct, emotion, intellect, will and so forth with reference to the adolescent, and with reference to the difference between the adolescent and the child. Considerable attention is paid to the aesthetic, moral, and religious aspects of the adolescent life. The book abounds in broad generalizations and one could wish for more specific information in their support. In many places one feels the lack of actual data. For example, when the author tells us with reference to the sense of smell in the adolescent that "the threshold is lowered, and the just observable difference in odors is very small," one feels the need of experimental evidence in confirmation of such a statement, and the same is true of many other similar statements found in the book.

R. PINTNER.

¹ Tracy, F.: *The Psychology of Adolescence*. Macmillan, 1921. Pp. 246.

A Critical Discussion of Project Methods.—Stevenson¹ has written a very able summary and criticism of existing concepts and practices of project teaching together with conservative formulation of his own. The project is defined as “a problematic act carried to completion in its natural setting.” The criteria are: (1) Information is to be acquired by reasoning rather than by memory; (2) information is to be acquired for its use in modifying conduct rather than for its own sake; (3) principles are to be introduced as they are needed in the solution of a problem rather than before the solution is begun; and (4) learning is to take place in a natural rather than in an artificial setting. The author emphasizes the last criterion; “The provision for the natural setting of the teaching situation is the distinctive contribution of the project method. Without the natural setting there is no project.”

Far from recommending a complete discard of the present curriculum, the author does not attempt to organize any subject completely on the project basis. He does believe that “at least certain units of the elementary and high school subjects” may be taught by projects. He believes in the scientific determination of minimum essentials and in seeing to it that they are learned. The projects are used so far as is wise or possible. “If it is found difficult to provide projects for these facts, or if the project method seems to be uneconomical, then the problem method or the method of presenting the material systematically should be utilized.” Always, there should be drill and review until “a systematic grasp of the subject is realized.” The project thus becomes a supplement to other methods rather than a complete substitute for them. They “help bridge the gap between school tasks and tasks carried on outside the school.”

The summaries and criticisms of definitions proposed by leading writers on project methods and the bunched pages of sample projects in many subjects, add greatly to the usefulness of the book.

A. I. G.

A Psychology for Laymen by an English Writer.—The author states that the *Psychology of Everyday Life*² is not an elementary textbook of psychology, nor a “popular account of some of the marvels of psychology with all the psychology left out,” but “the main facts of the

¹Stevenson, John Alford. *The Project Method of Teaching*. New York: Macmillan Co., 1921. Pp. XVI + 305.

²Drever, James. *The Psychology of Everyday Life*. New York: E. P. Dutton & Co., 1921. Pp. IX, + 164.

science so far as these touch the life of the man in the street." What we find is a series of short essays on such topics as instincts and emotions, imitation, suggestion, play, sensations, perception, memory, hallucinations and spiritualism, which include many definitions, descriptions and classifications of concepts. The book represents an effort to combine many of the connections of James, McDougall and Freud. In dealing with appetites, instincts, moods, emotions, sentiments and social organizations, the writings of McDougall are followed, largely; while in the chapters on perception, memory and thinking, the influence of James appears. In almost every chapter, the Freudian explanations are offered, and on the whole the writer is favorably inclined toward them. The book is interesting, but sometimes (as in dealing with emotions, moods and sentiments) rather confusing. Almost no space is given to the recent work of psychologists in mental testing or in professional and vocational fields other than psycho analysis.

A. I. G.

Sex Education for Boys.—This little book¹ gives a hundred pages of sound and useful advice to fathers concerning the enlightenment of boys in matters of sex. The author does not recommend sermons or punishments, but gives a series of "projects" by which the right information is provided at the opportune time, in a manner less sanctimonious and pedagogically more sound than is often found in books of this type.

A. I. G.

II. BRIEF NOTICES OF NEW GENERAL EDUCATIONAL BOOKS

1. RICHARDSON, M. W. *Making a High School Program*. School Efficiency Monograph Series. World Book Co., Yonkers, N. Y., 1921. Pp. VIII + 27. Paper.

A little manual telling in detail how to make a high school program on the block system, based on the writer's years of experience in making the program of the Girls' High School of Boston. Supt. F. V. Thompson reports that the plan is in operation in a number of Boston high schools and works well. All necessary forms, charts, diagrams, and illustrative programs are printed in the manual.

¹ Galloway, T. W. *The Father and His Boy*. New York; Association Press, 1921. Pp. XI + 199.

2. DAVIS, E. E. *The Twentieth Century Rural School*. Bobbs-Merrill, Indianapolis, 1921. 242 pages.

A well-written handbook for rural teachers. It abounds in concrete episodes that show examples of good and bad educational work by actual rural teachers and supervisors. All phases of rural school work are covered: the first approach, school library; getting the school before the people; some vitalizing educational agencies and organizations; school playgrounds; the social factor in rural life; making better citizens; salaries; school taxes in country districts; roads and communication; the public school and health of the country; rural school museum; a "standard" school; layer school units in the country.

3. LULL, H. G. and WILSON, H. B. *The Redirection of High School Instruction*. J. B. Lippincott, Philadelphia, 1921. 286 pages.

A brief text-book for reading circles and courses in methods in secondary education, which includes discussions of the administration of the curriculum, the administration of the student activities, and the selection and evaluation of subject matter. The book represents the theory of "a minimal essentials" course—"the social core of the curriculum" is constantly stressed—to be required of all pupils. It gives concrete illustrations for each major high school subject, of programs, courses, and devices. A definite discussion of the organization of instruction on the "project-problem" basis is given. A number of actual community and school surveys are reported.

4. FOSTER, H. H. *Principles of Teaching in Secondary Education*. New York: Scribner's Sons. 1921. Pp. XVIII 367.

This is a book in teaching methods for prospective or untrained teachers. It discusses such psychological matters as: instincts; interests and teaching; attention and teaching; associative learning; transfer of efficiency. It sets forth: different current aims of instruction; current practices in the conduct of the class exercise; the use of the question in teaching; recitations and lesson development: how lessons can employ problem-solving methods and secure appreciation and expression; standards and measurements in teaching.

5. MONROE, W. S. *Report of Bureau of Educational Research, Division of Educational Tests*. For 1919-1920. Urbana, Ill. University of Illinois, 1921. Paper. 64 pages. 25 cents.

A report of grade norms for 12 different standard tests in arithmetic, reading, language, history, algebra, geometry, and for general intelligence. Results of giving tests to pupils of Illinois schools are reported in the form of (1) grade medians, (2) number of pupils attaining indicated scores in each grade, (3) tables of percentile scores. A very valuable handbook compilation for users of standard tests in city schools.

6. DOUGLASS, H. R. *The Deriation and Standardization of a Series of Diagnostic Tests for the Fundamentals of First Year Algebra*. Eugene, Oregon: University of Oregon, 1921. Paper. Pp. 48.

Report of detailed investigation to determine what constitutes the fundamentals of first year algebra, and to devise a series of tests for testing ability and diagnosing weaknesses. Contains very good evaluation of Rugg-Clark and Hotz tests. Supplies suggested new tests.

THE JOURNAL OF EDUCATIONAL PSYCHOLOGY

Vol. XII

November, 1921

No. 8

IS THE RATING OF HUMAN CHARACTER PRACTICABLE?

HAROLD RUGG

The Lincoln School of Teachers College, and
Teachers College, Columbia University

CAN HUMAN CHARACTER BE "RATED" ON POINT SCALES ACCU- RATELY ENOUGH FOR PRACTICAL USES IN EDUCATION?

YES, AND—NO

Yes,—if the rating is done under conditions as rigorous as the following:

First, if each final rating given a person is the average of *three independent* ratings, each one made on a scale as objectified as the man-to-man-comparison type of scale.

Second, if the scales on which the ratings are made are comparable and equivalent,¹ having been made in conferences under the instruction of one skilled in rating scale work.

Third, if the three raters are so thoroughly acquainted with the person rated that they are competent to rate.

But these conditions are practically unattainable in public schools. Hence the answer to our original question—No, not by methods so far generally employed, and probably not unless methods of rerating and checking judgments are carried far beyond present practical possibilities.

We can now predict that, on a scale of 100 points, the probable error of the best single rating that we can get under "experimental" conditions is between 5 and 6 points. Thus a large proportion of even "experimental" ratings will locate persons outside his true "fifth" of the entire scale. And I assume that to locate a person within his proper "fifth" of the entire scale is a sound and practical working

¹ "Comparable and equivalent" explained in detail later in the article.

criterion for rating. Furthermore, we can predict that a single rating by a typical school officer—supervisor, superintendent, principal—will only rarely locate a person within his proper “fifth” of the entire scale. And we can predict that a single rating on any one of the commonly used types of scales,—like Elliott’s or Boyce’s or Beatty’s or Hill’s—will have a probable error of at least 10 points on a scale of 100 points and hence be practically valueless.

Hence the apparently dogmatic answer to the question—“Can human character be “rated” on point scales accurately enough for practical uses in Education”—No. We would far better give our energies to the attempt to measure it *objectively*, than to make *subjective* judgments of it *on point scales*. The point cannot be made too emphatically that we should discard these loose methods of rating once and for all. We cannot justify wasting the time of our school administrators and deluding our teachers with fictitious “ratings” and “marks.” Even on one of the so-called “standardized” point rating schemes a *single* rating has little or no scientific validity.

I propose to publish in these articles detailed evidence to aid us in setting straight our thinking about this matter. This evidence was collected under rare conditions—conditions that we may never be able to duplicate—certainly not unless we have another great war.

Let us refresh our memories a bit about educational rating scales. The movement is about eleven years old. Elliott made the first suggestion in 1910 with a very elaborate scheme of some hundred traits to which were assigned “weights” or “credits.” A group of traits like “dynamic efficiency” was given 80 points and each of several traits contributing to it was likewise assigned points—5 or 25, as the case might be. In this way a person was rated on a scale of 100 points. The weighting of the separate elements was entirely arbitrary.

I was in a group of about a dozen trained school officers in 1911 which used the Elliott scale. We rated the same group of ten teachers, sometimes several observers observing a teacher simultaneously. Of several hundred correlations made up from such ratings, practically no correlations exceeded 0.2. Many were negative. The conclusion was clear that a subjective rating scheme, made up of an elaborate set of abbreviated descriptions of traits, with weights arbitrarily assigned, and with rating done against no external standard, had little or no scientific validity. Yet the Elliott scale did much good in stimulating thought along the line of “How shall we measure these intangible and dynamic qualities?”

The Boyce scheme was little better,—forty-five “qualities,” described very briefly and with indefinite captions, rating in 10 divisions but against no common external standard. Rating was thoroughly subjective,—nothing external to the rater’s consciousness against which to measure the thing rated. Comparisons of independent ratings on this scheme showed insufficient agreement. (Boyce’s own data were so meagre as to be inconclusive.)

Other attempts were made but always of the same subjective type.

The first innovation and one promise of real progress was the man-to-man-comparison scale. This was the product of a seminar discussion at the Carnegie Institute of Technology—a class conducted by Professor Walter Dill Scott. The suggestion was applied in the development of rating scales for employees in industry. Here at last, it was said, was a method of judgment which took the process of rating out of the realm of the “subjective” and gave it scientific standing—made rating “objective.” The other scales had employed no common standard. The man judged was rated *against* nothing external. *Measurement*—ran the dictum,—*implied comparison with a scale*. The conclusion followed—make your “rating scale” a scale of human beings. The increments—the unit distances on your scale—will be the distances or differences between “scale men.” How select these scale men? First, by choosing “the best man you ever knew” and writing his name down at the top point on your rating scale. Second, by selecting “the poorest man you ever knew” and using him at the foot of the scale. Similarly, you chose an “average-man,” a “better-than-average-man,” and a “poorer-than-average-man,” making a five “point” scale. You gave numbers to the five men, like 15, 12, 9, 6, and 3, and your “scale” was complete.

Rating was simple on it. You simply compared your man with these five men. Was he as good as the best man,—say, Richardson? No. Was he better than Johnson, the poorest man? Yes,—by far. Was he superior to Bankson, the “average” man? Well, hardly. And so it went until he was finally placed on the scale and given a score.

The scheme was ingenious—a new and unique suggestion. It attracted attention from personnel managers in industry. But its bases and implications were not thought through. Its validity and reliability were not determined experimentally.

Just at this point, 1917, America entered the war, and the educational and psychological brains of the country were harnessed into two remarkable working teams—the Psychological Division of the Army

under Professors Yerkes, Terman and others, and the Committee on Classification of Personnel under Professors Scott, Bingham, Strong, Coss, and others.

Two important attempts to measure human abilities were made by the respective teams—the Army A, Alpha and Beta—group intelligence tests by the Psychological Division; and the Army Rating Scale for the rating of officers' efficiency by the Committee on Classification of Personnel.

The rating scale introduced was the man-to-man-comparison scale to which I have referred. Careful canvass of all the available scales was convincing of the fact that this scale was the most objective and gave promise of the greatest validity and reliability. *But practically no scientific evidence was at hand from which to establish its true reliability. Neither were tested methods of constructing and using scales.* So it was necessary to introduce it into the army without this scientific and basic evidence. Great credit should be given to Colonel Scott and his colleagues for their skill and patience in the face of opposition in finally securing a trial of this important suggested method of judging character in Army officers.

The use of the Army Rating Scale during the winter, spring and summer of 1918, raised grave questions in the Committee as to its validity and reliability. Preliminary experiments with it in certain camps tended to confirm the suspicion of its unreliability.

On being brought into the service of the Committee on Classification of Personnel, as Statistician, in September, 1918, I was commissioned to make an intensive analysis of the construction, use and reliability of the Army Rating Scale. For nearly three months this occupied the time of a staff of six statistical and clerical workers under my direction.¹

Conditions were set up under the stress of a great war that it would be difficult if not totally impossible to duplicate in peace times. The remarkable conditions under which this investigation was carried on cannot be stressed too strongly. I doubt if we shall ever have again—certainly not unless we are again thrown into a great war—the opportunity to duplicate them.

Imagine an experimental situation in which groups, each of nearly 100 very intelligent officers (a total of 461, with an average alpha

¹ Great credit should be given to Miss Cecile Colloton, my present research assistant, for her intelligent and thorough work in charge of the statistical treatment of the data.

score of B⁺), who had lived together for 11 to 14 months in depot brigade; who had associated constantly,—slept, dined, drilled, worked and played together, until they knew each other's personal characteristics as only can very intelligent, observant men who are literally bound together. Conditions would have to be set up as ideal as those of an Arctic exploring expedition to provide a better situation for the experiment. I say, imagine the remarkable circumstances in which whole groups of such men could be brought together willingly, gladly, for three day conferences on the rating scale; in which with meticulous care they would laboriously and immediately under supervision construct rating scales by our most refined technique; in which they would *use each other* for scale men on their scales; in which they would *rate each other* on these scales, and finally, signing their own names, give us their scales for scientific comparison.

Certainly no public educational situation in America can duplicate in our generation, experimental conditions so favorable to the construction and critique of rating scales. And the critical work was done on what is clearly the most objectified of all the scales suggested to the present time. Hence the importance of the data presented in this report.

EVIDENCE FROM THE INVESTIGATION OF THE ARMY RATING SCALE

The evidence for these introductory comments will be presented systematically. I reproduce first three samples of the man-to-man-comparison type of scale: A. The Army Rating Scale; B. The Rugg Rating Scale for Students; C. The Rugg Rating Scale for Teachers.

The problem before us is this: How closely is a single rating of a person's qualities made on the man-to-man scale a true measure of his qualities? Throughout the remainder of this article and the next one, the data and discussion will refer altogether to the Army Scale and the rating of the abilities of officers in the United States Army in 1917 and 1918. Analogies and applications to education will be drawn constantly.

Selection of Criteria for Judging Validity of Ratings.—The most difficult and important task we encountered at the outset of the investigation was the selection of criteria against which to measure the validity of ratings of character. Four were finally selected:

The determination of:

1. The degree to which a number of officers agree in rating the same officer independently, both in total rating and on specific contributory traits.

A. THE ARMY RATING SCALE

I. PHYSICAL QUALITIES.	
Physique, bearing, neatness, voice, energy, endurance.	Highest..... 15
Consider how he impresses his command in these respects.	High..... 12
	Middle..... 9
	Low..... 6
	Lowest..... 3
II. INTELLIGENCE.	
Accuracy, ease in learning; ability to grasp quickly the point of view of commanding officer, to issue clear and intelligent orders, to estimate a new situation, and to arrive at a sensible decision in a crisis.	Highest..... 15
	High..... 12
	Middle..... 9
	Low..... 6
	Lowest..... 3
III. LEADERSHIP.	
Initiative, force, self reliance, decisiveness, tact, ability to inspire men and to command their obedience, loyalty and co- operation.	Highest..... 15
	High..... 12
	Middle..... 9
	Low..... 6
	Lowest..... 3
IV. PERSONAL QUALITIES.	
Industry, dependability, loyalty; readiness to shoulder responsi- bility for his own acts; freedom from conceit and selfishness; readiness and ability to co- operate.	Highest..... 15
	High..... 12
	Middle..... 9
	Low..... 6
	Lowest..... 3
V. GENERAL VALUE TO THE SERVICE.	
Professional knowledge, skill and experience; success as adminis- trator and instructor; ability to get results.	Highest..... 40
	High..... 32
	Middle..... 24
	Low..... 16
	Lowest..... 8

B. THE RUGG RATING SCALE FOR HIGH SCHOOL STUDENTS

The Rating scale containing the names of typical students who can be compared with the student to be rated

(Primarily for teachers and principals to give the students a numerical rating)

I. Ability to learn—to assimilate new ideas*

Best student.....	38
Better than average.....	30
Average.....	22
Poorer than average.....	14
Poorest student.....	6

Summary numerical rating.....

II. Qualities of industry and attitude toward school work

Best student.....	38
Better than average.....	30
Average.....	22
Poorer than average.....	14
Poorest student.....	6

Summary numerical rating.....

III. Qualities of leadership

Best student.....	38
Better than average.....	30
Average.....	22
Poorer than average.....	14
Poorest student.....	6

Summary numerical rating.....

IV. Team-work qualities

Best student.....	38
Better than average.....	30
Average.....	22
Poorer than average.....	14
Poorest student.....	6

Summary numerical rating.....

V. Personal and social qualities

Best student.....	38
Better than average.....	30
Average.....	22
Poorer than average.....	14
Poorest student.....	6

Summary numerical rating.....

Total numerical rating.....

*Some forty odd questions are answered about the pupil's traits. These serve as complete definitions of the qualities. Copies of the two Rugg Rating Scales can be secured from the University of Chicago Bookstore, 5802 Ellis Ave., Chicago.

2. The degree to which officers' scales are comparable and represent equivalent amounts of the traits in question—personal qualities, physical qualities, intelligence, leadership, and general value to the service. (In place of these read for analogy to education the separate qualities on my scales.)

3. The degree to which the scale positions of officers used on the "Intelligence" element of the Rating Scale correspond to scale positions determined by three objective psychological tests.

4. The degree to which the Rating Scale detects differences in ability which are detected by other conspicuous measures of success. The most important was: *appointment to a captaincy from civil life without having had previous military experience.*

Four types of data were secured in the investigation:

1. More than 100,000 "official" quarterly ratings. (Once every three months each officer from colonel down in each army unit was rated on the Rating Scale by his immediate superior officer. The spring, summer and autumn ratings, 1918, were treated independently.)

2. Ratings obtained in September, 1918, at an officers' personnel school, conducted at Fort Sheridan by Lieutenant Colonel J. J. Coss.

3. Scales, ratings, re-ratings, and detailed personal data from the two groups of officers, 461 in all, who cooperated in an *experimental study of the rating scale* at Camp Sheridan and Camp Zachary Taylor.

4. Correlation and other data, of ratings and of other measures of success in the army, such as previous annual earnings, promotions, years of schooling, age and scores made upon psychological and alertness tests.

The rating of officers in the Army parallels in its practical features the rating of school teachers and students. It is important to stress the great emphasis on the *practical* needs of rating and to caution the reader that refinement in rating was not expected. The one criterion that was constantly uppermost in our minds was: Is the probability great that a rating given an officer by the use of the Army Rating Scale will locate him in that fifth of the entire scale in which he would be placed by an objective measure of success, if such could be found? Thus, it was assumed that the Army wished to discriminate officer ability with no greater degree of refinement than that which would classify all officers in five groups. This likewise is typical of our school situations. For convenience we may think of these as falling into the following numerical intervals on the scale, and as being

described by some such suggested code symbols as: A = 84-100, B = 68-83.9, C = 52-67.9, D = 36-51.9, E = 20-35.9.

Thus in our analysis we will keep constantly in mind that *approximate* accuracy is all that is demanded or that could be secured, under actual working conditions, either in the army or in the public schools.

Assumptions Implicit in the Man-to-man-comparison Scale.—The construction and use of the Army Rating Scale rested upon three fundamental assumptions:

1. That the scales made by various officers will be comparable and equivalent with respect to the absolute amounts of trait represented upon them. More concretely Captain Evans' scale for "Intelligence" will be comparable to Captain Brown's Scale for "Intelligence" in that the two "15" or "highest" men will represent approximately the same degree of the trait; similarly for the two "3" or lowest men, the "12" men, the "9" men and the "6" men. It should be pointed out that this assumption will be implicit in the construction of *any* rating scale, the purpose of which is to lead to an objective and absolute measure of an officer.

2. But this in turn points to the second assumption; namely, that officers of varying grades of ability are so distributed throughout the various units of the army (in camps, cantonments, schools, etc.) that *each rating officer will have represented on his original list, and on his scale approximately the same differentiation and distribution of the trait in question.* More concretely, that the spread and absolute amounts of the "physical qualities," say, on one officer's scale will be approximately equivalent to the spread and amounts on another officer's scale and that the intervals on the scale represent approximately the same differences in amount of the trait.

3. Finally, the assumption is implicit that army officers can be trained to evaluate the abilities of their associates and subordinates sufficiently to construct scales which will be comparable, and to make the "man-to-man-comparison" required for the rating of their subordinates.

Objective Measures of Success.—As in education, so in the army, very few adequate measures of success were available. For example, we had no measure of success in overseas fighting. To make possible a complete analysis of the validity of the Rating Scale a complete study should have been made of the relation between success in active service and all other facts which were available on officers: (1) their ratings on the rating scale, obtained both during the period of train-

ing and in overseas service; (2) their citations and special rewards of merit or demerit; (3) their promotions; (4) their previous annual earnings in civil life, years of schooling, age, etc.; (5) the degree to which they had exercised responsible control of men and of policies in previous civil life activities; (6) their preference for a particular branch of service and their success in it; etc. Had the war continued it is likely that this study would have been made.

We finally isolated four objective measures of success: 1. The first was found in the abilities of men conspicuously selected to officer the army. Officers were appointed from training camps to various commissions, some to the rank of 2d lieutenant; some to that of 1st lieutenant; and still others at once to captaincies. *It was confidently assumed that the men who were appointed to captaincies at once from civil life, without having had previous military experience, combined in an outstanding way the qualities demanded for success in the Army.* Thus we implied that any measuring device used on officers ought to measure men in the long run in approximately the same way that appointment to a captaincy from civil life measures them. Throughout this study, this criterion was regarded as an important one in checking up the measuring power of the Rating Scale. Furthermore it contributed to an analysis of other measures of success in the army.

2. The second objective measure of success in an officer is to be found in his *achievements* in carefully conducted *psychological tests*. During the previous six months evidence had accumulated rapidly that the army psychological tests measured quite closely the types of ability demanded for success in officers in the army. Reports issued about that time by the Division of Psychology, *e.g.* (1) "Army Mental Tests," (2) Reports numbers 26, 27 and 28 of the Psychological Board at Camp Wadsworth, South Carolina, offered a number of concrete illustrations of this fact.

In correlating achievements on the army psychological tests with ratings made upon the Army Rating Scale, it was recognized that the two instruments do not measure the same identical group of abilities. Lack of correlation is expected because of the fact that some of the qualities included in the Rating Scale (even in the "Intelligence" part of the Rating Scale) do not coincide with those involved in performance on the psychological test. Expressed in statistical terms we may say that there should be a reasonably "high correlation" between psychological test scores and ratings, or any other measure of success. A "high correlation," represented conservatively by an

"r" of 0.5 to 0.6, means that there is a very distinct tendency for differences in ability to be similarly detected by the two instruments.

3. As a third measure of success we canvassed the relation between ratings, and the typical qualification facts which were available on officers; their previous annual earnings; their appointment to commissions of various grades, their ratings, their promotions and their previous occupational activities. It was recognized that at best, promotions and "earnings" could be regarded as only partial measures of success in the army. Promotions, for example, were contributed to by so many factors other than that of military merit, that the "promotion interval" from say, appointment to "2d Lieutenancy" and subsequently to a captaincy must be regarded only as a very coarse measure of achievement. These measures were regarded only as supplementary.

4. With the foregoing measures of effectiveness of the Rating Scale there was followed the practical criterion that independent ratings to be valid measures of officers should show a very limited amount of variability. This criterion led to the chief statistical method of treating the data—the determination of the variability of independent ratings on officers by different raters.

HOW CLOSELY DO INDEPENDENT RATINGS OF CHARACTER AGREE?

Several elaborate sets of data were collected to answer this question:

1. The "official" quarterly ratings, over 100,000 in number.
2. The ratings of 325 men in an officers personnel school.
3. The "experimental" scales and ratings of each other obtained from 461 officers.

1. *The Study of Two or More Official Ratings as an Officer.*—The first step was to compare two and more official ratings made on an officer, both by the same rater, and by different raters. 2383 cases were tabulated; (the spring ratings had already been proven of little value so the comparison was made for summer and autumn ratings). For these average differences between the two ratings (on a total scale of 80 points it should be remembered) A—by the same rater were:

For Second Lieutenants.....	10.2 points
For First Lieutenants.....	10.2 points
For Captains.....	8.4 points

B. For different raters, the average differences between the two ratings were:

For Second Lieutenants.....	12.0 points
For First Lieutenants.....	21.7 points
For Captains.....	16.9 points

About half of the differences were increases and half decreases. The medians were somewhat different, but the general conclusion was inescapable: *it was very improbable that an officer was located within even his proper "fifth" of the entire scale by an "official" rating.* The ratings were practically valueless. Either the Rating Scale was being improperly used or else the task of constructing the scale and of making the man-to-man comparisons that are necessary is too difficult and complicated to be compassed under the practical limitations of army rating. (And these rating conditions are quite comparable, if not superior, to those of education—certainly as to education and experience of raters, administrative control over rating and the like.)

Evidence was accumulating that the scale was being improperly used. In fact it was known that at first and second quarterly rating, thousands of officers were "rated" without the use of a scale at all. We had evidence to show, however, that vast improvements were made in the October ratings. The net result of the study of these "official" ratings was to make clear the need for the study of agreements in rating a person when the whole procedure of constructing and using scales is definitely controlled.

I would caution the reader that in only a limited sense do the sweeping conclusions stated at the beginning of this article rest upon the analysis of these "official" ratings. They are impressive only as illustrations of the results that would likely follow from the utilization of such a rating scale (as this 1918 official scale was) in most school systems today. But my insistence that a *single rating of character on any judgment scale* is invalid and of little practical value rests upon far better evidence than was obtained from the study of these official ratings. We turn next to data distinctly better than those studied so far.

2. *Differences in 6 to 31 independent ratings on the same officer by associates in an officer personnel School.* I quote directly from Professor Coss's account of the way he secured the data of this part of our investigation:

"325 college men averaging about $20\frac{1}{2}$ years of age, who had been in training from 2 to 3 months in what was practically an Officer's Training Camp, were given a course in the work of the Personnel Adjutant. These men were rated by the Rating Scale and were given an intelligence rating.

"The officers who had instructed this group in military branches were called together for an evening meeting in a small ballroom of a nearby hotel and were instructed in the use of the scale. The instruction was given under particularly favorable circumstances, the attention was excellent, and the appreciation of the points to be noted seemed general. These officers then individually made a rating scale using second lieutenants as a base. They were then asked to get together by groups on the basis of the companies to which they had been attached. They were given lists of the men from their companies and were asked to rate each man. They worked with interest and carefully. The Company groups turned in ratings both from the individual members and from the average of their grades.

"One of the exercises of the school was the study of the Rating Scale. The scale was explained. The soldiers were then required to read from the printed matter on the scale and made a scale using second lieutenants as a base. They then rated each other. Each man from every company rated all the other members of that company and turned in his rating sheet. Captain Trabue of the Surgeon General's office conducted the intelligence test examination in which the men took a lively interest. The intelligence rating of the group was extraordinarily high."

Table I supplies the findings for the first 15 men in company I. This is a thoroughly random sample of all the data. The typical range of ratings on an officer, 30 to 25, shows that any one rating selected at random may be a very unreliable estimate of an officer's true rating. He would be displaced by two whole—even three whole—divisions of the scale. Some of his associates called him "excellent" officer material, while others rated him as distinctly poor. We had many instances in which an officer was rated in the top fifth by one associate and in the bottom fifth by another.

Note the striking constancy of the standard deviations around 8 and of the probable error between 5 and 6. The conclusion is clear that for probabilities of 20 to 1 we are assured that an officer will be

TABLE I.—AVERAGES AND MEASURES OF VARIABILITY OF 6-31 INDEPENDENT RATINGS ON 15 OFFICERS IN A PERSONNEL SCHOOL AT FORT SHERIDAN
(These 15 are typical of entire group of 325)

No. of officer rated	No. of ratings on him	Range of ratings on him	His average rating	Average deviation of ratings on him	Standard deviation	Probable error
1	27	52-80	65.7	6.1	8.42	5.67
2	23	38-67	52.9	6.7	8.11	5.47
3	27	66-92	80.9	6.4	7.61	5.13
4	30	36-73	53.5	6.4	8.50	5.73
5	19	53-81	63.8	5.4	7.10	4.79
6	31	48-83	64.4	5.8	7.52	5.07
7	31	43-77	62.2	4.6	6.23	4.20
8	28	43-71	56.6	5.9	7.16	4.83
9	23	39-75	55.2	7.7	10.00	6.74
10	27	32-65	48.3	6.4	8.52	5.75
11	27	46-74	59.4	6.6	7.83	5.28
12	29	48-89	75.1	8.2	10.34	6.97
13	20	37-70	54.0	5.9	7.46	5.03
14	25	37-66	54.1	6.2	7.73	5.21
15	25	43-82	61.9	7.3	9.23	6.23

correctly placed only within a range of about 15 to 18 points, more than one "fifth" of the scale itself.

(To be continued.)

(In the December installment will be printed the data secured in the experimental studies at Camps Sheridan and Taylor, and part of the analysis of the psychological factors involved.)

SUBJECTIVE TESTS VS. OBJECTIVE TESTS

A. A. ROBÁCK

Harvard University

It would seem more of an anachronism for any psychologist to advocate at this advanced stage of the testing game the introduction of tests which require a certain amount of interpretative ability on the part of the examinee and considerable judgment on the part of the tester. Mental testing was, of course, bound in the direction of an objective goal for various reasons, chiefly, however, on account of the needed uniformity in the scoring which would involve possibly millions of cases. Absolute standardization of the scoring directions and computations was the desideratum of the intelligence test movement. Very rarely a subjective test would spring up such as the interpretation of fables and the description of pictures; but in general, intelligence tests grew more machine-like or call it objective, if you like, from year to year, until the "multiple choice" sort of test has become most popular both with examiners and examinees, and with good cause; for both had an easier task before them under such conditions than when the answer was not suggested on the examination. The value of such tests was indeed manifest at the time of the war when the stupendous number of men examined would have made it quite impossible to carry out the examinations with the dispatch and efficiency shown by the corps of intelligence testers engaged. In drawing the line between a feeble-minded person and one of normal intelligence such tests will doubtless furnish us with considerable diagnostic information, but it is questionable whether a method where the correct answers are supplied together with several possible but incorrect answers could give us an insight into the mental caliber of the individual examined or allow for the numerous variations to be expected in a large group.

Objective tests are satisfactory only in mathematical and mechanical problems, where only one solution is possible; and even there, one may long to take into consideration the different modes of approach discoverable in a group. In many other objective tests, however, there are decided disadvantages, and mental testers are only deceiving themselves when they suppose that the merit of a test is to be adjudged on the basis of the scorer's time and energy it saves. It is undeniably true that those tests in which a number of possibilities are put before the individual who is to under score or appose a cross to the correct answer—tests which I should designate by the name of "multiple

choice" problems—are the most economical, but when we stop to analyse the situation as the examinees are confronted with the problem, can we assert with any measure of certainty that they are thinking out the solution? In tests of this kind it seems to me we are dealing with factors which may be regarded as components of intelligence, but which certainly cannot present themselves as the characteristic marks of mentality. To analyze a hypothetical instance, suppose we ask our examinee to underscore one of the following reasons for going to school:

1. to get an education;
2. to earn more money later on;
3. to provide teachers with work;
4. to have a good time;
5. to become useful citizens.

It is evident that a bright individual who would respond to a natural query of this sort in an off-hand, yet correct manner, might become distracted by, if not bewildered at, the absurd possibilities offered and in consequence fumble about before making his final reaction. It would appear that the more direct and original a person is, the more apt would he be to flounder. The mediocre person in this case gains an advantage over the superior intellect. Other factors that count here are (a) suggestibility, (b) motor-co-ordination in manipulating the pencil, and (c) rapidity of decision. The *intellectual*, though more intelligent than the "red-blood" would take a longer time to decide between two alternatives either of which something might be said in favor of, whereas he might have readily arrived at a sound conclusion, if the alternatives were not suggested to him.

We should perhaps be willing to admit that ease of decision, non-suggestibility, control of inhibitions and even motor co-ordination all enter into the make-up of intelligence, but I for one should be chary about viewing them as representative factors. I can well conceive of a superior mind with slow reactions or of a clever man who yet is suggestible. The mere fact that one person can underscore a number of words a trifle sooner than another does not make the former more intelligent.

There is a widespread tendency among intelligence testers to produce the pragmatic criterion in meeting the contentions of critics, and one may anticipate in this connection an argument such as this, "What matters it whether our subjects belong to one type or another? It is the results that count, and since A has a score of 200 while B

managed to obtain only 198, we may safely conclude that A is the more intelligent of the two."

In reply to this form of reasoning, I should say that unless we define *intelligence as a certain score which one receives in a particular set of tests*, we are not warranted in assuming that a higher score is absolutely indicative of higher intelligence and a lower score of lower intelligence. It still remains to be discovered, for instance, just how much of a given degree of success is due to temperamental traits and volitional tendencies of one sort or another, not to mention, of course, the part played by experience.

When we stop to examine the source from which the "multiple choice" tests derive their objectivity, we shall probably trace it to the arbitrariness of the deviser. Since he cannot exhaust the list of all the possible answers to a given problem, he must content himself with the supplementation of the few which occur to him as approximating in some way the correct solution. If instead of approximating the requisite answer, the few possibilities suggested are remote, such as *hat: head: glove: boat: lamp: table: hand: key*, the test must fall short of the purpose for which it was intended, unless it was designed to mark off the feeble-minded from the normal.

On the other hand, if the other possibilities suggested are just about as correct as the requisite answer, *e.g.* in the case of the analogy *square: triangle: circle: ellipse: cone: semi-circle: arc: oval*, we have a subjective situation to deal with; for although the term *cone* seems to complete the most appropriate analogy, there is something to be said for each of the other forms as satisfying the requirement. The degree of *subjectivity or objectivity*, then, which attaches to any test of the "multiple choice" kind would depend on the parity or disparity of the possibilities supplied. But let us be mindful of the fact that in real life, alternatives that present themselves for action resemble the former category, hence the selective process called forth by the test is in no way representative of choice in actual life.

A further objection against these cut-and-dried objective tests is the *lack of provision for such qualities as initiative to figure in the examination*. The man who can beat out a new track is given no more credit than the one who happens to choose the right path by noting that the others are only blind alleys.

But the most serious difficulty of all that we have to contend with in our "multiple choice" tests is the *comparatively small scope they afford us to tap the higher mental capacities which serve to distinguish*

the superior intellect from the average mind. The testing of reasoning has been confined to the purely logico-mathematical problems; and, what is worse, has been conducted wholly along the "true-false" method. Just how much of the score in that particular type of test is due to guess-work and how much to ratiocination is a matter not easily ascertainable.

Tests of abstraction, interpretation, tests to determine the degree of acumen or subtlety have been utterly neglected. The same applies to tests for critical ability, expression, and judgment tests, the sore need of which has occasioned the contrivance of substitutes which measure some phase of intelligence, but just what phase is not at all clear. Fancy the critic in actual life who is always provided by benevolent people with the cue of his criticism. Or what should we think of the interpreter of new movements and phenomena who must needs have at his elbow an inventory of possible interpretations out of which he selects the proper one? It is with such considerations before me that, in preparing my tests for superior adults, I was led to deviate from the highway of dubious efficiency and resort to the seemingly narrow by-way of precision, requiring an *answer* and not a line or a cross. The scoring is thereby rendered more difficult, but to compensate for the additional expenditure of time and effort we may repose far greater confidence in our results. If the old principle that we take out of an enterprise just about as much as we put into it holds true anywhere, it is in the field of mental testing; and any sacrifice of accuracy at the shrine of speed is deplorable on general grounds.

The element of subjectivity in scoring tests in which the answers are not supplied will naturally be present to some extent, but experience has taught me that when the scoring directions are definite and complete and the examiners are asked to abide rigidly by the standards, the amount of personal bias entering into the scoring would be reduced to a negligible quantity. To be sure, it would require some training and more than a modicum of intelligence to score a series of tests for superior adults, but once the directions are followed and each test or portion of a test is scored by one person, uniformity is insured.

In reply to critics who would urge that any test which does not make use of the device of objective scoring is bound to involve a personal factor, I should maintain that even if such fear be well-founded, there is no alternative under the circumstances unless we are prepared to delude ourselves into the belief that we are testing

superior intelligence when in reality what we are doing is tapping various degrees of mediocrity, and that too in a rough way.

The multiple choice method is incontestably a good device in experimenting on animals, but when applied to men and women of high intelligence it affords us no adequate measure of the qualities we are bent on testing. After all, is it not the examinee's intelligence that we are concerned with instead of the arduous labor of the examiner? Is it not more in keeping with scientific procedure to test a smaller number or fewer groups painstakingly rather than to accumulate a vast body of data that are not wholly reliable because of the insidious infiltration of vitiating factors? One must realize in setting forth this argument, that under certain conditions "half a loaf is better than nothing" and we should accordingly be guided by our purpose and the circumstances in a given case. Thus during the great War, it was not the purpose of the intelligence examination committee to study individual differences or to single out men with exceptional mentality. Nor could anything but an objective set of tests be manipulated considering the huge size of the army to undergo the examination.

In testing adults, however, for superior intelligence we have before us a different situation. In the first place, the number examined will necessarily be limited. The person responsible for the scoring would be expected to make himself thoroughly conversant with the units of measurement and the points of each problem, but should not be required to overtax his faculty of judgment, that is to say, the nature of the test should be such as to demand concise and clear-cut answers.

That the difficulty to be encountered in scoring "unguided" tests is exaggerated had become evident to me when I finished marking some 300 acumen tests which at first seemed like one of the Herculean labors. It was probably the most subjective of my whole series,¹ and it looked as if the variety of answers and modes of tackling the questions would baffle me; but I soon learnt to discriminate between the correct and incorrect answers, with the result that, as far as I was concerned, the scoring appeared to be essentially the same as in the "true-false" tests. To take one or two instances: when the examinee is asked to show the significance of the adjectives in the following two clauses

- (a) He thought it would be a *difficult but interesting task*,
- (b) but it proved to be an *interesting but difficult task*.

¹ Roback: "Mentality Tests for Superior Adults."

Only one type of answer is admissible, and that is that in (a) the idea foremost in the mind was the interesting nature of the work, but in (b) the difficulty of the task outweighed the enjoyment derived from the interest. Other modes of expression might be equally suitable but the point of "interest foremost in (a) and difficulty foremost in (b)" would have to be stated. Similarly whatever else might be said about the connotative difference between the words "*haste and speed*," the correct answer should relate the former to the *agent's mental state* and the latter to the objective result of *covering a lot of space in a relatively short time*.

SUMMARY

Our objections to a set of tests which is made up exclusively of the cut-and-dried kind, either on the multiple choice plan or along the "true-false" line, may be enumerated as follows:

I. The objectivity of the tests does not attach to the general method of procedure, taking in the whole situation—purpose of the examination, mental functions of the examinee, etc.,—but is confined to the scoring only. In other words, the tests are devised with a view to the ease of scoring. Objectivity, such as this, is at bottom illusory, for its very *raison-d'être* is the subjective desire of saving time and labor.

II. The superior adult not only misses the opportunity for manifesting his ability under such conditions, but his very originality and initiative in thought become a burden to him, when the courses are mapped out for him, with the result that the mediocre person has the advantage over his intellectual superior.

III. Purely objective tests must necessarily be artificial, in no way representing a life situation.

IV. Some, at least, of the higher functions cannot be approached by objective tests. Interpretation, analysis, subtlety, power of expression, judgment and other abilities are inaccessible to the "multiple choice" or "true-false" tester.

V. The factor of guess-work in a given test of that kind is indeterminate. In close scores, unless the disturbing element is eliminated, we have no means for proper comparison.

VI. Objective tests afford us no avenue to the study of individual differences; and if differential psychology plays an important part anywhere in the different levels of intelligence, it is obviously in those levels above the average.

AN EXPERIMENTAL AND STATISTICAL STUDY OF READING AND READING TESTS (Concluded)

ARTHUR I. GATES

Teachers College, Columbia University

MONROE'S STANDARDIZED SILENT READING TEST¹

This test is a revision of the Kansas Silent Reading Test, devised by F. J. Kelly in 1916.² It consists of from 14 to 16 short paragraphs each followed by a question which is answered by writing or underlining a word. The rate of reading is determined by giving a credit for each paragraph read regardless of the answer. The comprehension score is the sum of the values of those paragraphs the questions of which are correctly answered. Five minutes time is allowed. The test includes a fore-exercise of one paragraph. Test I is designed for grades III, IV, V; Test II for grades VI, VII, and VIII.

A defect of some importance in Monroe's test is the fact that the number of successful responses that can be secured by chance, is rather large. In the case of several paragraphs the chance of succeeding is one in two. The scale has been tested by marking the reaction without reading the paragraph. In the case of Test II, Form 1, a comprehension score of 12 was secured by checking the answers without reading the paragraphs. This score represents a grade of ability equal to the beginning of grade IV.

Forms I and II were given at intervals of nearly three months. The following correlations obtain between the two performances:

	Grade IV	Grade V	Grade VI
Rate Score.....	0.72	0.73	0.80
Comp. Score.....	0.70	0.50	0.56

These correlations give us little direct information concerning the tests except the suggestion that the rate scores are more stable than the comprehension scores, possibly for the reason that the factor of chance success in the latter is rather great.

¹ Monroe, W. S.: Monroe's Standardized Silent Reading Test. *Journal of Educational Psychology*, 1918-19, 9, 303-312.

² Kelly, F. J.: Kansas Silent Reading Tests. *Journal of Educational Psychology*, 1916-17, 7, 63-80.

TABLE VIII.—CORRELATIONS OF THORNDIKE-McCALL READING TEST WITH—

Grade	Stan- ford men- tal age	Comp. group tests	Brown rate	Brown comp.	Courtis rate	Courtis comp.	Mon- roe rate	Mon- roe comp.	Bur- gess	Direc- tions	Special vocab.	Holley vocab.	Gray's oral	Comp. comp.	Correct comp. comp.	Comp. rate
III	0.16	0.60	0.62	0.02	0.74	0.75	0.60	0.78	0.79	0.80	0.83	0.88	0.87	0.82	0.88	0.83
IV	0.35	0.74	0.24	0.37	0.23	0.31	0.52	0.42	0.34	0.39	0.38	0.56	0.54	0.69	0.76	0.43
V	0.62	0.84	0.46	0.07	-0.03	0.35	0.86	0.48	0.56	0.52	0.40	0.52	0.26	0.74	0.75	0.45
VI	0.75	0.74	0.34	0.25	0.39	0.42	0.41	0.45	0.38	0.53	0.42	0.66	0.63	0.63	0.64	0.42
VII	...	0.59	0.30	-0.24	0.33	0.53	0.29	0.03	0.18	0.37	0.52	0.50
VIII	...	0.62	0.17	0.50	0.29	0.40	0.52	0.59	0.71	0.81	0.50
Mean.....	0.47	0.69	0.36	0.16	0.33	0.46	0.60	0.53	0.48	0.45	0.51	0.56	0.57	0.66	0.73	0.52
S. D.	0.19	0.09	0.14	0.24	0.23	0.15	0.16	0.14	0.17	0.25	0.19	0.21	0.22	0.14	0.12	0.14

We have found the test to be too easy for the upper grades. Seventy per cent of our VIII grade pupils and over fifty per cent of the VII grade obtain a perfect rate score on Test 2 which is designed for grades VI, VII, and VIII. It was therefore impossible to use these results for purposes of correlation.

In our opinion, the Monroe test would be a more useful instrument if it were constructed on the principle of the Thorndike-McCall, by combining Tests 1 and 2 with an extension into more difficult material than the most difficult Test 1. As the tests are now constructed, Test 2 duplicates (in terms of difficulty) fifty per cent of the material contained in Test 1 and includes only one paragraph of greater difficulty. There are many practical advantages for having a continuity of norms from grades III to VIII. With the Monroe test as it is, the brightest children in grade V cannot be tested by Test 1 and of those who can, no comparisons can be conveniently made with performances in the upper grades.

Monroe has secured, it seems, a somewhat better method of controlling comprehension while rate is being measured than either Brown or Courtis but his method is, in one respect, distinctly different from that used by Burgess. In the rate score Monroe gives credit even if the child fails to comprehend; Burgess gives no credit in such a case. Using the technique of partial correlations, Pressy¹ found that Monroe's rate score, comprehension eliminated—or rather held constant—yields a slightly negative correlation with teachers' estimates of reading ability. The validity of his teachers' judgments as criteria is, of course, unknown. His correlations (0.27 for rate and 0.38 for comprehension) are very low when compared with those found in the present study.

Table XII gives the correlations of the Monroe rate and comprehension scores with all other measures. "Rate" yields the same correlations with both composites as "comprehension" but both appear to agree more closely with the composite for rate. The correlations between Monroe's rate and comprehension scores average $0.92 \pm S. D. 0.03$. Both scores correlate highly with Burgess and Directions but not so well with Thorndike-McCall and Courtis comprehension. From our studies of many cases known to be of extraordinary slowness but of good understanding² it was clear that the Monroe test measures either a quite different type of comprehension or

¹ *School and Society*, 1920, 11, p. 746.

² One such case is described in the discussion of the Thorndike-McCall.

TABLE XII.—SHOWING CORRELATIONS OF MONROE'S READING TEST, RATE SCORE, WITH—

Grade	Stan- ford mental age	Comp. group intell.	Brown rate	Brown comp.	Courtis rate	Courtis comp.	Mon- roe rate	Monroe comp.	Bur- gess	Thorn- dike- McCall	Direc- tions	Special vocab.	Holley vocab.	Gray's oral	Comp. comp.	Correct comp.	Comp. rate
III	0.21	0.21	0.76	-0.33	0.65	0.51	0.90	0.61	0.60	0.65	0.60	0.51	0.62	0.58	0.62	0.87
IV	-0.06	0.28	0.74	0.28	0.48	0.48	0.93	0.70	0.52	0.70	0.58	0.60	0.61	0.82	0.81	0.82
V	0.30	0.24	0.72	-0.34	0.38	0.10	0.96	0.74	0.86	0.68	0.59	0.42	0.60	0.72	0.78	0.88
VI	0.52	0.79	0.48	0.07	0.74	0.83	0.88	0.83	0.41	0.84	0.72	0.46	0.29	0.86	0.84	0.91
Mean..	0.24	0.38	0.68	-0.08	0.56	0.48	0.92	0.72	0.60	0.72	0.62	0.50	0.53	0.75	0.76	0.87
S. D. . .	0.21	0.24	0.11	0.27	0.11	0.26	0.03	0.08	0.16	0.07	0.06	0.09	0.14	0.11	0.10	0.03

MONROE'S COMPREHENSION SCORE WITH—

Grade	Stan- ford mental age	Comp. group intell.	Brown rate	Brown comp.	Courtis rate	Courtis comp.	Mon- roe rate	Monroe comp.	Bur- gess	Thorn- dike- McCall	Direc- tions	Special vocab.	Holley vocab.	Gray's oral	Comp. comp.	Correct comp.	Comp. rate
III	0.31	0.40	0.67	-0.25	0.60	0.51	0.90	0.70	0.78	0.70	0.78	0.66	0.75	0.66	0.72	0.87
IV	-0.08	0.35	0.63	0.23	0.42	0.39	0.93	0.66	0.42	0.70	0.50	0.60	0.65	0.70	0.74	0.76
V	0.70	0.22	0.68	-0.26	0.39	0.03	0.96	0.66	0.48	0.67	0.62	0.38	0.63	0.69	0.72	0.85
VI	0.56	0.78	0.60	-0.02	0.73	0.76	0.88	0.80	0.45	0.80	0.43	0.43	0.43	0.91	0.90	0.86
Mean..	0.37	0.44	0.65	-0.08	0.54	0.42	0.92	0.71	0.53	0.72	0.58	0.52	0.62	0.74	0.77	0.83
S. D. . .	0.30	0.21	0.03	0.20	0.14	0.26	0.03	0.07	0.14	0.05	0.13	0.11	0.11	0.10	0.08	0.04

else measures it in a quite different way from the Thorndike-McCall. The slow readers attain a low score on the former and a high score on the latter. If a measure of power of comprehension freed of the mechanical factors in reading is desired, the Thorndike-McCall is the test to use.

The Monroe test yields only fair correlations with intelligence tests but higher with group tests than with the Stanford-Binet. The mean correlations with vocabulary tests range from 0.52 to 0.72. The correlations with Gray's test of oral mechanics are fairly high (0.53 and 0.62). The Brown comprehension score shows a zero correlation as it does with most other tests. On the whole, no evidence appears in support of Pressy's finding that the Monroe comprehension score is more valid than the rate score; in fact, the rate score yields slightly better correlations with both speed and comprehension composites.

COURTIS SILENT READING TEST No. 2

The Courtis test measures speed and comprehension separately. Part I consists of a childish narrative of 567 words to be read for three minutes, the position being checked each half minute by encircling a word. In Part II the same material is broken into 14 paragraphs each followed by 5 questions which are to be answered in a word. The number of questions answered correctly in 5 minutes is the score for comprehension. The subject may reread a paragraph as often as desired. The author uses an "Index of Comprehension" which is the percentage of correct answers. We have found this percentage to be so frequently 100 in grades above the third that it has not been used in this study.

The Courtis rate test provides no check upon the degree of comprehension with which the subject reads. There is nothing to prevent radical changes in the speed on different tests save personal checks adopted by the subject. Following are correlations of a test with a re-test given three months later.

Grade III	0.85
IV.....	0.87
V	0.70
VI.....	0.48
VII.....	0.57
VIII.....	0.52
Mean.....	0.666
<i>S. D.</i>	0.16

While it is impossible to determine what allowances to make for the long intervals, it appears that the consistency of performance on this test is not as high as is desirable, especially in the upper grades, for which the material is rather trivial and uninteresting. Generally, a positive correlation between initial ability and improbability in a function is found and if such were the case here, the interval would make the correlation higher than it would have been after a day.

For Part II—the comprehension test—the correlations of the two trials were:

Grade III.....	0.80
IV.....	0.78
V.....	0.65
VI.....	0.80
VII.....	0.80
VIII.....	0.76
Mean.....	0.765
S. D.....	0.05

The constancy of performance is higher and less variable in this part of the test which controls comprehension. Our data cannot be considered as trustworthy evidence on the reliability of the tests since it is impossible to take into account the effect of the interval.

Table XIII shows the correlations of the mean scores of two Courtis tests with other single tests and the composites. The rate score yields a mean composite of $0.76 \pm \text{S. D. } 18$ with the composite of speed and $0.58 \pm \text{S.D. } 19$ with the corrected composite of comprehension. The comprehension score yields a higher correlation with the comprehension composite ($0.70 \pm \text{S. D. } 14$) and a lower correlation with speed $62 \pm \text{S. D. } 0.20$. The difference however is not reliable and if we make allowance for the inclusion of rate score in rate composite and the same for comprehension, it appears that either part of the test is about as good as the other and that neither tests one aspect of reading any better than the other, in spite of the fact that the correlation of Courtis speed and comprehension averages but 0.44. The range of the speed-comprehension correlations for Courtis is from -0.27 to $+0.82$, a fact to be explained partly, in our opinion, by the possibility of shifting the type of reading—now reading carefully, now rapidly—in the speed test. The comprehension score yields a higher correlation than the rate score with both the Brown and the Monroe rates, as well as with the Thorndike. The correlations with Burgess and Monroe comprehension are about the same.

TABLE XIII.—SHOWING THE CORRELATION OF THE COURTIS TEST, RATE SCORE, WITH—

Grade	Stan- ford men- tal age	Comp. group intell.	Brown rate	Brown comp.	Courtis rate	Courtis comp.	Mon- roe rate	Mon- roe comp.	Bur- gess	Thorn- dike- McCall	Direc- tions	Special vocab.	Holley vocab.	Gray's oral	Comp. comp.	Correct comp.	Comp. speed
III	0.05	0.25	0.87	-0.09	0.82	0.65	0.60	0.65	0.74	0.69	0.47	0.61	0.67	0.73	0.71	0.81
IV	-0.36	0.57	0.43	-0.03	0.15	0.48	0.42	0.69	0.23	0.39	0.32	0.30	0.53	0.49	0.54	0.73
V	0.17	-0.08	0.22	-0.60	-0.27	0.38	0.39	0.37	-0.03	0.41	0.03	0.47	0.13	0.15	0.23	0.39
VI	0.40	0.64	0.51	-0.12	0.66	0.74	0.73	0.71	0.39	0.78	0.54	0.45	0.60	0.75	0.78	0.80
VII	0.57	0.48	0.48	-0.02	0.70	0.77	0.33	0.32	0.37	0.40	0.49	0.87
VIII	0.18	0.65	0.05	0.58	0.51	0.29	0.47	0.46	0.73	0.96
Mean..	0.07	0.36	0.53	-0.13	0.44	0.56	0.62	0.33	0.52	0.34	0.45	0.48	0.50	0.58	0.76
S. D. . .	0.34	0.26	0.20	0.21	0.38	0.11	0.14	0.23	0.20	0.20	0.10	0.21	0.20	0.19	0.18

CORRELATIONS OF THE COURTIS TEST, COMPREHENSION SCORE, WITH—

	III	IV	V	VI	VII	VIII	Mean..	S. D. . .
III	-0.06	0.35	0.74	0.08	0.82	0.51	0.72
IV	-0.34	0.57	0.47	-0.17	0.15	0.48	0.38
V	-0.07	0.53	0.57	0.06	-0.27	0.10	0.39
VI	0.45	0.70	0.49	-0.05	0.66	0.83	0.24
VII	0.74	0.60	0.15	0.70	0.71
VIII	0.52	0.53	-0.20	0.58	0.68
Mean..	0.01	0.57	0.57	-0.02	0.44	0.48	0.54
S. D. . .	0.29	0.13	0.09	0.13	0.37	0.26	0.18

Evidence secured from our poor readers indicates that both forms of this test really measure speed rather than power of comprehension, after the manner of the Thorndike test. Our Grade VIII poor reader is distinctly the slowest performer in the class in this test but above the average in Thorndike-McCall. The second form of the test seems to be in general more reliable since it is less easy to "fake" it.

The correlations with Stanford-Binet mental age average zero, grade IV being negative and grade VI positive. The correlations with the composite of group intelligence tests are rather low, averaging $0.35 \pm \text{S. D. } 0.26$ for Part I and $0.57 \pm \text{S. D. } 0.13$ for Part II. Correlations with the vocabulary tests are low. In general, the S.D.s for the Courtis tests are very high, indicating either great variability of performance or individual differences in type of reading, some of them of the sort not indicative of real ability as shown by other tests. The correlations with other criteria are not as high as those yielded by Burgess and Monroe.

WOODWORTH-WELLS DIRECTIONS TEST

The Directions used were the three forms devised by Woodworth and Wells¹ combined into one test. Those authors had in mind testing ability to understand instructions. To quote their words: "The conditions which it was sought to meet in the test material are: (1) that the motor response should be very simple and quickly performed; (2) that the instructions should be very simple, but varied; and (3) that the instructions should be as concise as possible in order that reading time might not be the determining factor." Samples of the easy directions are: "Cross out the *g* in *tiger*." "Put a dot in the circle below the center *O*." There are 40 directions of this or slightly greater difficulty² and 20 are considerably more difficult. For example: "Write *yes*, no matter whether China is in Africa or not. . . ; and then give a wrong answer to this question: "How many days are there in the week? . . ."

The results were scored by the familiar method—right minus the sum of errors and omissions. Three and a half minutes were allowed

¹ Woodworth, R. S. and Wells, F. L.: Association Tests. *Psychol. Monograph* No. 57, Dec., 1911.

² Pintner, R. and Toops, H., have empirically determined the difficulty of these directions and published a revision. See *Journal of Educational Psychol.*, 1918, 9, 123-142. We used the old forms because the material was needed in a hurry and electrotypes were at hand.

but many of the lower grade subjects did not get as far as the Hard Directions. Woodworth found, however, a correlation of 0.92 between the two forms, so they appear to yield approximately the same results. Table XIV gives the correlations.

The Directions test seems to be more like the Burgess than any other here used. This similarity is evidenced by the fact that it yields a higher mean correlation with the Burgess than with any other. In certain respects the Directions test is superior—in reducing the amount and difficulty of the motor response to a minimum. In some instances the Directions test put a demand on information (*e.g.* “If Edison discovered America,” etc.) and in other cases the directions are something of a verbal puzzle. A scale which combined the merits of the two tests would probably be superior to any test for rate of reading now available.

With the exception of grade IV the correlations are fairly high and the test agrees, like the Burgess, about as well with the composite of comprehension as with speed. On the whole, it appears to measure reading ability about as well as most of the newer instruments. The correlation with Thorndike-McCall is low but, like the latter, it shows an increased correlation with Stanford-Binet as we pass from the lower grades up. With the exception of grade VII, correlations with the group intelligence tests are 0.6 or above and correlations with the vocabulary tests are around 0.5.

GRAY'S ORAL READING TEST¹

This test consists of eleven paragraphs arranged in order of increasing difficulty. It requires some skill to use the tests in accordance with Gray's directions. The time for reading each paragraph is taken with a stop watch and the number of errors noted. A table of credits representing a composite of speed and errors is provided and the result is multiplied by a figure to make allowances for grade differences, *i.e.*, the higher the grade the lower the credit allowed. All of this is time-consuming and the result obtained is a score which allows comparisons only with norms and scores in that particular grade. This method may be pedagogically sound but in practical use it is very cumbersome, and teachers become chagrined at the time involved. In this study we have taken simply the sum of the credits for the paragraphs from

¹ Gray, W. S.: *Studies of Elementary School Reading Through Standardized Tests*. Univ. Chicago *Sup. Educ. Monograph*, Vol. 1, No. 1.

TABLE XIV.—SHOWING THE CORRELATIONS OF THE DIRECTIONS TEST WITH—

Grade	Stan- ford men- tal age	Comp. group intell.	Brown rate	Brown comp.	Courtis rate	Courtis comp.	Mon- roe rate	Mon- roe comp.	Bur- dick- gess	Thorn- dike- McCall	Special vocab.	Holley vocab.	Gray's ora	Comp. comp.	Correct comp. comp.	Comp. rate
III	0.18	0.66	0.70	-0.05	0.69	0.75	0.65	0.70	0.76	0.80	0.69	0.80	0.76	0.89	0.91	0.81
IV	0.12	0.65	0.68	0.12	0.39	0.28	0.70	0.70	0.80	0.39	0.60	0.52	0.65	0.55	0.58	0.54
V	0.48	0.59	0.72	-0.23	0.41	0.27	0.68	0.67	0.84	0.52	0.63	0.27	0.67	0.84	0.89	0.81
VI	0.61	0.82	0.56	-0.01	0.78	0.78	0.84	0.80	0.87	0.53	0.54	0.62	0.48	0.90	0.90	0.90
VII	0.31	0.51	0.26	0.32	0.37	0.55	0.03	0.47	0.60	0.60
Mean.....	0.35	0.61	0.63	0.02	0.52	0.49	0.72	0.72	0.76	0.45	0.62	0.46	0.64	0.73	0.78	0.79
S. D.	0.20	0.17	0.08	0.16	0.20	0.23	0.07	0.05	0.11	0.25	0.04	0.24	0.10	0.15	0.14	0.14

TABLE XV.—SHOWING CORRELATION OF GRAY'S ORAL READING TEST WITH—

Grade	Stan- ford men- tal age	Comp. group intell.	Brown rate	Brown comp.	Courtis rate	Courtis comp.	Mon- roe rate	Mon- roe comp.	Bur- dick- gess	Thorn- dike- McCall	Direc- tions	Special vocab.	Holley vocab.	Comp. comp.	Correct comp. comp.	Comp. rate
III	0.13	0.56	0.57	0.07	0.67	0.72	0.62	0.75	0.74	0.87	0.76	0.63	0.13	0.84	0.82	0.78
IV	0.09	0.42	0.54	-0.03	0.53	0.58	0.61	0.65	0.58	0.54	0.65	0.53	0.40	0.52	0.50	0.38
V	0.16	0.18	0.61	-0.07	0.13	0.21	0.60	0.63	0.48	0.26	0.67	0.40	0.04	0.51	0.54	0.69
VI	0.35	0.52	0.47	0.06	0.60	0.45	0.29	0.43	0.39	0.63	0.48	0.48	0.51	0.51	0.50	0.42
Mean.....	0.19	0.42	0.55	0.02	0.48	0.49	0.53	0.62	0.55	0.58	0.64	0.51	0.27	0.60	0.59	0.57
S. D.	0.08	0.15	0.05	0.04	0.21	0.19	0.14	0.11	0.13	0.22	0.10	0.08	0.19	0.14	0.13	0.17

Gray's table of composite speed and error scores, making all scores comparable.

A system of shorthand practices is provided by Gray who suggests checking the child's reading for six types of errors; gross errors, minor errors, omissions, substitutions, insertions and repetitions. While it requires some practice and skill to do this, the detailed information is frequently most valuable. The subjects must be treated individually, of course, with this test which requires from five to ten minutes.

We have been unable to give re-tests with this instrument. Unfortunately but one form is available. No objective measure of comprehension is provided, since the test deals frankly with the mechanics of oral reading.

Table XV gives the results. The test yields a correlation of 0.59 with the criterion of comprehension and 0.57 with rate. The correlations with individual and group measures of intelligence are lower than those obtained by the best tests of silent reading. Correlations appear of approximately 0.5 with a single test of silent reading or of vocabulary. It yields a correlation of about 0.5 with a pronunciation test consisting of 36 words, containing from two to thirteen letters, used by the writer for diagnostic work. The criteria presented in this paper do not enable us to test the real value of this test.

We have found it to be an exceedingly useful instrument, especially for purposes of individual diagnosis partly for the reason that the one who experiments can observe intimately the particular reactions. Its use for such purposes will be described in another paper. There is a real need for several new forms of this test.

THE VOCABULARY TESTS

Holley's Sentence Vocabulary Scale was devised for the purpose of measuring intelligence.¹ It consists of seventy words from the Stanford-Binet series printed in short sentences which are to be completed by underlining one of four words which follow each. For example: "Some puddles are made of . . . mud . . . sand . . . stone . . . brick." No time limit is fixed.

Before the writer became aware of the existence of the Holley Scale, a similar test had been devised consisting of fifty words from

¹ Holley, C. E.: *Mental Tests for School Use*. Bureau of Educ. Research, University of Illinois, *Bull.* No. 4, pp. 86-91.

the Terman list, each followed by five words, one of which was to be underlined to illustrate the meaning. For example:

1. Gown—(dress, tree, bird, ram, fish.)

44. perfunctory—(skillfully, odorous, speedily, carelessly, pretty.)

There was no time limit.

The Thorndike Visual Vocabulary was used in all grades, but an error in administration forced the results to be discarded except for grades V and VIII. In these grades the following correlations appeared:

	Thorndike visual vocabulary with							
	Mental age	Group intell.	Burgess	Thorndike-McCall	Holley	Special vocab.	Comp. comp.	Comp. speed
Grade V.....	0.30	0.53	0.56	0.50	0.48	0.53	0.56	0.48
Grade VIII.....	0.47	0.49	0.47	0.52	0.56	0.58	0.53

Tables XVI and XVII give the results for the other two tests. They show almost identical correlations throughout. The correlations with rate and with comprehension in reading range from 0.4 to 0.6. The correlation with group tests of intelligence is 0.5 and about 0.3 with the Stanford-Binet mental age.

In the mass, knowledge of word meanings is positively associated with reading ability but the correlation is not sufficiently high to make a vocabulary test an adequate measure of it. The correlation of vocabulary and the composite of comprehension is 0.6 as compared to 0.8 for the Burgess test. In special cases we find wide variations between performance in reading and vocabulary tests. Likewise the vocabulary tests here used are positively—but not very closely—related to intelligence. Our knowledge of the functions involved in acquisition of word meanings is very meagre and hazy. We do not know to what degree various types of training—specific training with words, wide reading, etc.—will increase vocabulary. Certain studies suggest that general intelligence may set a rather strict limit upon such development. There is need for a review of the scattered data concerning vocabularies and a great need for extensive experimental and statistical study.

TABLE XVI.—SHOWING THE CORRELATIONS OF HOLLEY'S SENTENCE VOCABULARY SCALE 3A WITH—

Grade	Stan- ford men- tal age	Comp. group intell.	Brown rate	Brown comp.	Courtis rate	Courtis comp.	Mon- roe rate	Mon- roe comp.	Bur- dick- gess McCall	Direct- tions	Special vocab.	Gray's oral	Comp. comp.	Correct comp. comp.	Comp. rate
III	0.32	0.69	0.54	-0.03	0.61	0.67	0.51	0.66	0.82	0.80	0.77	0.13	0.82	0.80	0.77
IV	0.20	0.45	0.45	0.05	0.30	0.35	0.60	0.60	0.51	0.56	0.52	0.40	0.70	0.72	0.65
V	0.12	0.49	0.41	-0.34	0.47	0.42	0.42	0.38	0.48	0.52	0.57	0.04	0.48	0.56	0.43
VI	0.65	0.60	0.31	0.16	0.45	0.44	0.46	0.43	0.48	0.66	0.69	0.51	0.58	0.57	0.45
VII	0.61	0.65	0.13	0.37	0.22	0.46	0.18	0.07	0.28	0.47	0.48
VIII	0.18	0.18	0.24	0.47	0.21	0.51	0.59	0.52	0.51	0.45
Mean.....	0.32	0.50	0.42	0.03	0.45	0.35	0.50	0.52	0.54	0.46	0.69	0.27	0.56	0.61	0.54
S. D.	0.20	0.15	0.15	0.22	0.10	0.21	0.09	0.11	0.12	0.24	0.07	0.19	0.17	0.12	0.13

TABLE XVII.—SHOWING CORRELATIONS OF THE SPECIAL VOCABULARY TEST WITH—

Grade	Stan- ford men- tal age	Comp. group intell.	Brown rate	Brown comp.	Courtis rate	Courtis comp.	Mon- roe rate	Mon- roe comp.	Bur- dick- gess McCall	Direct- tions	Holley vocab.	Gray's oral	Comp. comp.	Correct comp. comp.	Comp. rate
III	0.26	0.58	0.53	-0.08	0.47	0.46	0.60	0.78	0.83	0.69	0.77	0.63	0.66	0.68	0.66
IV	0.19	0.54	0.48	0.08	0.32	0.40	0.58	0.50	0.38	0.60	0.72	0.53	0.64	0.67	0.72
V	0.44	0.31	0.60	-0.33	0.03	0.04	0.59	0.62	0.54	0.63	0.57	0.40	0.53	0.55	0.65
VI	0.49	0.52	0.32	0.04	0.54	0.36	0.72	0.43	0.74	0.54	0.69	0.48	0.58	0.59	0.52
Mean.....	0.35	0.49	0.48	-0.07	0.34	0.32	0.62	0.58	0.51	0.62	0.69	0.51	0.60	0.62	0.64
S. D.	0.12	0.10	0.10	0.16	0.20	0.16	0.06	0.13	0.19	0.05	0.07	0.08	0.05	0.05	0.07

The completion of a most extensive inventory of words used in English reading by Thorndike¹ offers new possibilities for scale construction and research in this field. More than 4,500,000 words from a selected list of sources were tabulated and of these, the 10,000 occurring most frequently have been printed in alphabetical form with indexes indicating the relative frequency of occurrence.² These words should form the content of tests for purposes of standardizing age and grade achievements, for diagnostic and experimental work, as well as for many other educational uses.

CORRELATIONS WITH AGE, INTELLIGENCE AND PERFORMANCE IN OTHER SCHOOL FUNCTIONS

The criteria were as follows:

1. Chronological Age,
2. Stanford-Binet Mental Age,
3. A composite of 6 to 8 group tests of intelligence,
4. A composite of three spelling tests including 182 words selected from several columns of the Ayres-Buckingham list,
5. A composite of the Woody Arithmetic tests (all four operations), Monroe's Diagnostic (12 to 24 functions), and Monroe's Reasoning Test,
6. A composite of speed and quality in writing,
7. A composite of school achievement including reading, spelling, and arithmetic,
8. Judgments by 4 to 9 teachers of "school attitude" consisting of what each teacher thought important such as application, diligence, persistence, interest, willingness, etc.

Table XVIII gives the correlations with comprehension and Table XIX with rate.

Both rate and comprehension are negatively correlated with chronological age and positively related to mental age. This is the usual finding: the correlations with mental age are low for grade III and increase to 0.6 or 0.7 in grade VI. Since the inter-correlations among reading tests were as high in the lower grades as in the higher, the increasing correlation with mental age may be interpreted to mean

¹ Thorndike, E. L.: Word Knowledge in the Elementary School. *Teachers College Record*, Sept., 1921, pp. 334-370.

² Thorndike, E. L.: "The Teacher's Word Book." New York: Teachers College Bureau of Publications, 1921.

that intelligence, as measured by Stanford-Binet shows itself only when the mechanics of reading are fairly well mastered. Other interpretations are possible and none can be wholly justified by our data.

TABLE XVIII.—CORRELATIONS OF READING COMPREHENSION WITH—

	Chron. age	Men- tal age	Comp. group intell.	Spell.	Arith.	Comp. achieve.	Writ- ing	School atti- tude
Grade 3.....	-0.16	0.10	0.65	0.63	0.17	0.77	0.51	0.46
4.....	0.05	0.16	0.77	0.54	0.28	0.72	0.41	0.22
5.....	0.16	0.41	0.68	0.04	0.11	0.81	-0.36	0.29
6.....	-0.63	0.69	0.88	0.36	0.26	0.71	-0.06	0.43
7.....	-0.39	0.58	0.25	0.19	0.70	0.07	0.32
8.....	-0.33	0.59	0.57	0.41	0.72	0.00	-0.22
Mean.....	-0.22	0.34	0.69	0.40	0.24	0.74	0.10	0.25
S. D.....	0.27	0.24	0.10	0.21	0.10	0.04	0.29	0.23

TABLE XIX.—CORRELATIONS OF READING RATE WITH—

	Chron. age	Men- tal age	Group intell.	Spell.	Arith.	Comp. achieve.	Writ- ing	School atti- tude
Grade 3.....	-0.30	0.19	0.43	0.52	0.21	0.79	0.31	0.55
4.....	-0.19	0.13	0.44	0.22	0.21	0.36	0.54	-0.32
5.....	-0.23	0.56	0.39	0.32	0.22	0.82	-0.46	-0.13
6.....	-0.49	0.60	0.76	0.38	0.36	0.74	0.00	0.45
7.....	-0.50	0.69	0.42	0.29	0.84	0.16	0.54
8.....	-0.39	0.73	0.42	0.15	0.69	0.19	-0.40
Mean.....	-0.35	0.31	0.57	0.38	0.24	0.71	0.12	0.12
S. D.....	0.12	0.30	0.15	0.09	0.07	0.16	0.30	0.40

The correlations with the composite of group intelligence tests is higher than with the Stanford-Binet and these are about as high in the lower as in the higher grades. Both of these facts might be explained by the greater demands of the group tests on reading, which are rather uniformly stable in the various grades, but this explanation is in no way defensible by our data.

The correlations with school attitudes are ambiguous. The significance of the teachers' judgments as indicated by the correlations,

are not uniform. Whether different judges are estimating different traits—or the same traits with varied success, or whether these traits are quite differently distributed with reference to reading ability in the different grades are possibilities which the data do not disclose.

The correlations of reading with the composite of school performance are high (0.70) and rather uniformly so among the grades. The correlations with single composite subjects are not high; the correlation of 0.4 with spelling being the highest, following by arithmetic which is 0.24. Writing shows, in the mean, a correlation of approximately zero but it is worth noting that the correlation is irregular with some indication of being positive in the lowest grades.

The measures of comprehension and rate correlate about equally with the several criteria since they are themselves highly correlated.

It must be noted again that the correlations of Tables XVIII and XIX are valid only for comparison among themselves. They do not represent ideal relations of these functions among unselected groups. They do, however, clearly indicate a rather low degree of inter-dependence of school functions which has an important bearing upon the validity, for example, of practices based upon the concept of the "accomplishment quotient" and other devices which assume but slight specialization of abilities.

GENERAL SUMMARY AND CONCLUSIONS

I. Concerning Reading Ability in General.

1. The concept of "silent reading ability" is justified both for theoretical and practical purposes by our data.
 - (a) A single comprehension test given in $3\frac{1}{2}$ to 30 minutes yields a correlation of 0.7 to 0.8 with a composite of comprehension tests representing from 4 to 8 hours of reading.
 - (b) A single rate test given in 1 to 5 minutes yields a correlation of 0.7 to 0.8 with a composite of rate tests.
 - (c) If corrections were made for attenuation and for the restriction of range in our data, the correlations would certainly be higher than they are.
2. Rate and comprehension are very highly correlated; the composites for the two yield an uncorrected correlation of $0.84 \pm \text{S. D. } 0.08$.¹

¹ The S. D. is the S. D. of the correlations for the separate grades from the mean of the grade correlations.

3. Most of the tests do not differentiate *rate* from *comprehension* for the correlations of rate tests with the composite of comprehension are about the same as with the composite of rate, and the correlations of comprehension tests are about the same with rate as with comprehension.
4. The existence was discovered in dealing with special cases of backwardness in the mechanics of reading of a useful distinction between rate and comprehension, of importance for diagnostic and remedial work.
5. The correlation of a single test with the composite averages 0.7 or higher, while the mean inter-correlation of single tests is about 0.5. While the former is higher partly because of less attenuation due to the fact that the composite is a more nearly perfect score, the difference is so great as to indicate that the several tests measure somewhat different combinations of the many functions involved in reading.
6. The correlation of silent reading with oral reading as represented by Gray's Oral Reading Test is nearly 0.6, which in our data is a fairly high correlation.
7. Correlations with two vocabulary tests (Holley's and one devised by the writer) average about 0.6.
8. Correlations with a composite of school subjects (reading, spelling, and arithmetic) are as high as 0.7 largely because *reading* is itself included in the composite.
9. Correlations of comprehension with spelling average 0.40, with arithmetic 0.24, with writing 0.10. The correlations of these subjects with rate are about the same.
10. Correlations with chronological age are negative; -0.22 for comprehension; -0.35 for rate.
11. Correlations with Stanford-Binet Mental Age are not high; 0.34 for comprehension and 0.31 for rate.
 - (a) The correlations of mental age with comprehension in grade III is 0.10 which rises steadily to 0.69 in grade VI. The facts are similar for rate. This is not to be explained by the lack of validity or reliability of the reading tests, or variability in reading performance because the correlations among reading tests are as high in the lower as in the higher grades. It is not likely that they are due to lack of validity of the Stanford-Binet tests. It is probable that intelligence of the type measured by the

Stanford-Binet does not show itself in reading until the mechanics of reading are fairly well mastered. There are, however, other possible explanations.

12. The composite of group intelligence tests yields higher correlations with reading than does the Stanford-Binet Mental Age; the mean with comprehension is 0.69 and with rate 0.64.

(a) The correlations in this case do not increase regularly as we advance from grade to grade.

13. Correlations with teachers' estimates of "school attitudes" such as interest, application, etc. are ambiguous, being high in some grades and low in others, for reasons which we have not discovered.

II. *Conclusions Concerning Particular Types of Reading Tests.*

A. *Tests for Rate of Reading.*

1. The mean of the grade correlations for rate tests with the criteria for rate was: Brown $0.82 \pm \text{S. D. } 0.18$; Monroe 0.87 ± 0.03 and Woodworth-Well's Directions $0.79 \pm \text{S. D. } 0.14$; Courtis 0.76 ± 0.18 .
2. Assuming the validity of the criterion, one of these tests appears to be as good as any other.
3. The criterion is, of course, not perfect and more detailed study of the reliability as well as the validity of the tests indicated.
 - (a) That a rate test which provides no control of comprehension (Courtis) gives less reliable results than one which provides a partial control by requiring a reproduction following the reading (Brown).
 - (b) That a test which mechanically controls comprehension by completing a picture, answering questions, carrying out directions, etc. is best (Burgess, Directions or Monroe).
4. The correlations of the rate tests with the criteria of comprehension are high and support the statements in the above I (3). They are: Courtis — $0.58 \pm \text{S. D. } 0.19$; Brown $0.66 \pm \text{S. D. } 0.10$; Burgess $0.80 \pm \text{S. D. } 0.09$; Directions 0.78 ± 0.14 ; and Monroe $0.76 \pm \text{S. D. } 0.10$.
5. Usually the so-called "rate" tests measure *comprehension* as well as *rate*, and the so-called "comprehension" tests measure *rate* as well as *comprehension*. It is well to use both parts of the combined tests (Courtis, Monroe)

since the results are more reliable but they do not measure perceptibly different functions of reading.

B. Tests of Comprehension.

1. The mean of the grade correlations with the criteria of comprehension are: Burgess $0.80 \pm$ S. D. 0.09; Brown comprehension $0.16 \pm$ S. D. 0.16; Courtis comprehension $0.70 \pm$ S. D. 0.14; Directions $0.78 \pm$ 0.14; Monroe comprehension $0.77 \pm$ S. D. 0.08; and Thorndike-McCall $0.73 \pm$ S. D. 0.12.
2. Assuming the validity of our criteria of comprehension, it appears that a record of what is written in a free reproduction following the reading of a passage (Brown test) is not a valid measure of comprehension.
3. Comprehension may be adequately measured by answering questions about short paragraphs which are of uniform difficulty (Courtis) or of increasing difficulty (Thorndike and Thorndike-McCall); by completing a picture in accordance with directions contained in short paragraphs of equal difficulty (Burgess) or by checking, crossing out, underlining or writing words to indicate understanding (Monroe and Woodworth-Wells Directions).
4. From the mean correlations with the criteria it appears that one test is about as good and measures about the same thing as another but by applying these tests to certain cases of backwardness in the mechanics of reading it was found.
 - (a) That the Thorndike and Thorndike-McCall measures power of comprehensions freed of the mechanical factors (speed) of reading.
 - (b) That success in the other tests is determined by speed as well as by comprehension.
 - (c) These facts are indicated by the data (5) below.
5. The correlations of comprehension tests with the criteria or rate are: Brown $0.08 \pm$ S. D. 0.25; Burgess $0.82 \pm$ S. D. 0.15; Courtis $0.62 \pm$ S. D. 0.20; Monroe $0.83 \pm$ S. D. 0.04; Directions $0.79 \pm$ S. D. 0.14 and Thorndike $0.52 \pm$ S. D. 0.14.

C. General Conclusions Concerning the Tests.

1. Detailed study of the individual tests betrays defects chiefly of the following types:

- (a) Too much time is spent in writing or drawing answers, solving linguistic puzzles and in other irrelevant functions.
 - (b) Certain variables, such as time spent in drawing, are not taken into account in standardizing the test units with the result that units are of unequal difficulty.
 - (c) The play of chance in carrying out directions is sometimes too great.
 - (d) The units are often too coarse or too few.
 - (e) Material is sometimes too easy or too difficult, trivial or uninteresting.
 - (f) Methods of scoring are unsatisfactory.
 - (g) Different forms of the test are of unequal difficulty.
2. Reading is a function which can be profitably measured and in which rate and comprehension can be differentiated, although most of our tests do not do so. The present tests are useful but not perfect instruments. We need tests constructed with such care that the numerous defects found in the tests now in existence shall be avoided.

APPENDIX

Table VIII, which gives the correlations for the Thorndike-McCall test was unintentionally omitted from the October issue. It appears in this issue.

SMALLER VS. LARGER UNITS IN LEARNING TO TYPEWRITE

J. W. BARTON

Professor of Psychology, University of Idaho

When one applies the general psychological principle that economy in human energy requires that for learning, "one should always begin by doing a thing as nearly as possible in the way it is eventually to be done,"¹ to typewriting, it is not surprising to find many instances of responses that do not conform to this general statement of requirements in adequate learning. There are many instances, in the lesson guides now in use, of exercises that are performed in ways about as far from the way in which they are eventually to be done as one could make them. It is likely that in many other fields of learning—penmanship, spelling, music, prerequisites in college and industrial work—there is to be found this same tendency to abandon this scientific principle of learning with great waste to the learner in time and energy required.

"Faculty psychology" seems to have made such decided inroads on our practices that it is hard to weed out in one or two generations the apparent transfer-of-training idea in the work of our schools. In some instances of learning it is very difficult to recognize that we are dealing with just this matter. Learning by means of too small parts is only another form of this "general-culture" notion; particularly is this the case if the parts are relatively much smaller than is required for comprehensive units, for if we should look upon learning as a means of fixing neural pathways of stimulus-response processes, then it is not difficult to understand why the "alphabet method" in reading was forced to give way to the larger unit methods now in use.

What has been found true for ineffective learning, in the case of reading is here found to be true for the alphabet-method in typewriting. The two situations seem very much alike and the results indicated below are very suggestive of what might be accomplished in the prevention of waste in the acquiring of this skill. What has been assumed to be true in learning—that however a bodily² process is exercised it will be equally efficient in any subsequent demands made upon it—is being shown, in many instances, to be fallacious and is thought to

¹ Hollingworth and Poffenberger: "Applied Psychology," 1917, pp. 66-67.

² It would be more nearly correct to use the term *mental* in this case except for the fact that it is now being recognized that the physical serves as a better means of approach in matters of control.

involve avoidable waste. This is particularly true in situations involving units too small to get all the benefits to be had from the readiness in response conditions and those of identity in neural factors involved on subsequent response occasions.

The Problem.—As far as could be determined at the time, typewriting is now being taught by the small unit, or alphabet, method in most institutions including this work. They all have in mind the two objects of accomplishment—mastery of key board and facility of operation—but the idea seems to be that mastery and facility in any situation is to effectively serve in others. Mr. J. W. Ross¹ points out that the letter combinations learned in one word are no particular help for a different combination of the same letters in another word. This conclusion is in keeping with the modern psychological view of partial identity in neural impulse responses. In criticising the word method Mr. Ross says,² “This means that in the word method the transition from writing the combinations of letters in one word to the combinations in another word requires a momentary pause in the thought process necessitated by the effort to break up a manual habit that was not directed by a conscious mental effort. This mental pause is eliminated in the line method which establishes an uninterrupted flow of mental direction coordinated with a corresponding smoothness in manual operation.”

It is likely a fact³ that there is a slowing up in such a situation due to lack of organization in the two different nerve processes involved in the two different word situations, or it could be explained on the basis of lack of “readiness”⁴ in nerve preparation for action. If such is the case in the situations indicated by Ross, what is to be said concerning such letter combinations as are found in his first exercise? These are, “asdf jkl; ;lkj fdsa jkl; ;lkj fdsa asdf jkl; ;lkj; ;fdsa.”

These are given, of course, for purposes of mastering the key-board; but if there is one bit of justification for the general principle of learning indicated above and for what Ross says for the units of his line method, it should follow that such exercises as those presented above are very inadequate as a means of teaching typewriting, since many of them never occur in English composition. Particularly has this conclusion some justification if it can be shown that the key-board can be mastered

¹ Ross: “Lessons in Touch Typewriting,” 1914, Preface.

² *Ibid.*

³ No scientific evidence available.

⁴ E. L. Thorndike: “Educational Psychology Briefer Course,” p. 53.

by means of exercises involving letter combinations had in the actual composition material most in keeping with the work of the typist.

What is held to be true of letter combination situations is also true for words or for sentences in so far as they fall short of the general principle indicated above, for learning in any situation of isolation has in it this lack of neural organization so necessary to subsequent responses if they are to be adequate. It is likely that Peterson¹ is very near the actual situation in learning when he defines it, partially at least, in terms of completeness of response. It might be well to point out that another factor here involved in learning, is that of "fitness in neural organization." It seems that for every new situation presented there is a more or less readjustment throughout. This readjustment likely consists, for the most part, of nerve units involved in the matter of making ready for the adequate response. To meet these demands of learning is to provide situations requiring a minimum of reorganization to the point of immediate adjustment, as well as to provide a maximum of exercise in the complete context of activity as it is to be used when learned.

It was this very apparent lack of fitness in the situations of response, involved in these isolated exercises so widely used in the teaching of typewriting, that prompted this study. It was felt that what had been true concerning the inadequacy of the alphabet method in teaching reading would likely prove true for the "isolated-small-unit-system" used in the teaching of typewriting, and that a better way might be had by conforming more closely to the psychologically justifiable larger unit exercise material plan.

The Method.—Two groups of students were selected, the first (fifteen subjects) by the process of the regular registration in high school and a second group (twelve) by special registration at a later date.

The first group began work the first Monday in September, 1918 under the direction of a teacher whose scholastic qualifications consisted of a high school training, two full years of college work, and a graduate of a standard business college. She was without previous teaching experience.

Nothing was said to the teacher at the time of registration that another class was to be started later, for which reason the students were directed according to what such teacher, just out of such training,

¹ *Psychological Review*, Vol. XXIII, No. 2, Mar., 1906.

thought to be the most approved methods, using the Remington Lessons as a guide.

During the eleventh week after the opening of school a part of the second group (eight subjects) was started. Later four additional ones were taken in. These students (known hereafter as group W and the regularly registered group shall be referred to as group P) were selected by advertising that another class in type-writing was to be started for a limited number of additional students. While no mental tests were used by which the relative mentality of the individuals was determined, the average class standings in all subjects for the year 1918-1919 indicates for the two groups a high uniformity in success at school work. The marks are as follows:

Group P		Group W	
Subjects	Grades	Subjects	Grades
1	B	1	C
2	B	2	C
3	B	3	C
4	C	4	B
5	A	5	B
6	C	5	B
7	B	7	B
8	C	8	A
9	C	9	C
10	B	10	D
11	B	11	C
12	C	12	C
13	D		
14	C		
15	B		

The two groups were directed by the same teacher, for the most part, and were under the same instructions in typewriting in every way except in the matter of the kind of exercise material worked upon and in a few other matters of handing in the product of their work at the office and reporting for a few other minor instructions.

Group P used the Remington Guide as the source of exercise material, taking the lessons in consecutive order as they are given there. The subjects of Group W were first asked to make a chart

of the keyboard on common letter size news print paper to be used in learning the keyboard. This chart, as well as the large one on the wall, was marked off by means of very distinct lines to indicate the respective fingers for the various sets of keys; then they were briefly instructed concerning the function of the various parts of the machine. At the end of these instructions the attention of the students was directed to the instructions written on the board. They read as follows: "Write a letter to Sears and Roebuck ordering a pair of shoes which should cost \$7.00 and to be sent by parcel post."

The next exercise included a letter to a friend inviting her to an evening party. In some cases, where it was found necessary, they were asked to continue to rewrite the letters indicated above. It should be understood that in all such cases the student was free to formulate anew these letters and that they were never required, at this point in the learning, to copy the letters already written.

This kind of self composed exercise material was used *until each student felt sure that she could properly produce* the desired characters without the aid of the key-board chart. At this point in the learning each student was required to do copy work.¹ The copy work consisted, for the most part, of reading matter found in the Red Cross materials of the time, although many of the students used their history texts or other school books as copy material, believing that by so doing they could get some help in the preparation of these lessons while doing the typewriting.

None of the students of either group had access to a machine outside of the forty minute period per day for five days per week throughout the year, except for the case of No. 11 in Group W who began much later than the other students of the group. It should be remembered that some of the students carried their keyboard charts home with them at first and used them in learning the work of the respective fingers in letter symbol production.²

In no case was it ever found necessary to check a student for attempting to drill on any isolated material for correcting letter errors. In all cases involving such errors, the students were required

¹ The writer appreciates that the material used in the composition or in the copy material does not conform to what has been found to be the most commonly used letter combinations in business or professional composition. A more scientific selection of such material should be had.

² One student (No. 8) memorized the keyboard during the first practice period, and during the evening at home by means of the chart.

to make the correction by reproducing enough of the context to include a unit of thought factor.

Data and Discussion,—It will be seen at once that for Group P it was impossible to get any measure of the product in terms of words per minute before January third. About all that can be done at present by way of standardizing such work is to require the exercises, as outlined in the guide in use, and demand that these be reproduced up to a certain degree of perfection in duplication. This was done, but it was of little service in getting a measure of the learning.

This class of work was participated in by Group P until about the middle of December (the students not being uniform in the amount of work accomplished an exact date in this matter is impossible) when they had reached that part of the guide that presents composition material. From this point on to the end of the course, the two groups were put together on the material of the guide.

After two weeks of practice on the composition material by Group P (Jan. 3, 1919) the following relative conditions obtained for the two groups:

	Group P	Group W
Average number of practice periods.....	75.9	20.75
Average number words attempted per minute.....	53.2	66.3
Average number of errors per minute.....	11.4	7.25
Average number words per practice period.....	0.70	3.19

On the basis of the time spent in practice, Group W produced 4.56 times as good results as was the case for Group P. The students of Group W, up to this time, spent only 0.273 as much time as those of Group P had spent and had attained 4.56 as great speed in production for the time spent as well as reducing the number of errors much below what was found for Group P.¹

From the third week in January up to the time of closing school in the spring (May 28) measures were taken, at intervals of from one to two weeks apart, of the number of words per minute written

¹ Nothing is claimed for the value of the composition material used for practice purposes in this experiment. Possibly even greater differences could have been had in the results of the two groups if material had been used that more nearly conforms to what is had by way of letter combinations required for business and professional composition; but even with the kind used, the results are very suggestive of what might be had with the more scientifically selected material.

by each subject. The figures, which immediately follow, represent words per minute after the exercise in each case had been penalized ten words for each error made.

GROUP P

Subjects	Words per minute for respective measures							Finals
	1	2	3	4	5	6	7	
1	27.4	31.0	29.2	28.6	37.4	40.8	39.9	49.4
2	48.3	45.8	49.8	47.4	49.6	52.6	49.33	55.2
3	14.8	22.22	16.2	17.6	21.7	24.0	24.8	26.3
4	22.6	25.6	23.4	18.2	25.8	27.4	30.8	37.1
5	44.1	49.0	43.2	41.0	47.0	47.2	42.0	56.8
6	33.4	36.6	37.4	25.2	28.6	35.6	30.8	39.4
7	34.9	35.6	38.4	36.8	37.4	45.6	35.6	46.6
8	33.3	29.8	29.8	24.2	17.2	37.2	31.1	43.0
9	13.3	15.5	21.4	17.8	13.8	13.8	16.2	20.6
10	28.9	36.2	31.0	28.4	32.1	36.2	33.1	39.5
11	33.1	33.6	37.4	25.0	33.1	45.0	38.8	40.6
12	30.2	21.4	21.4	20.2	33.9	29.7	23.7	43.2
13	20.9	27.2	23.2	17.6	21.3	25.4	30.1	34.6
14	22.1	31.2	23.1	23.2	26.3	24.2	28.8	30.3
15	34.9	45.2	37.2	32.0	31.7	34.0	41.4	51.8
Medians	30.2	31.2	29.8	25.2	32.1	35.6	33.1	40.6

GROUP W

Subjects	Words per minute for respective measures							Finals
	1	2	3	4	5	6	7	
1	18.2	22.0	18.6	19.6	18.5	25.4	30.0
2	39.6	33.0	35.0	41.7	41.2	49.4	48.00
3	20.4	17.0	22.0	23.8	31.4	24.8	34.4
4	This student would not use touch system							
5	50.4	69.4	45.0	60.1	56.0	66.0	55.9	70.2
6	28.7	24.8	30.6	32.5	29.0	32.9	43.9	37.2
7	32.6	32.0	32.4	29.0	32.9	40.6	57.4	46.0
8	45.3	52.2	52.0	48.6	45.2	61.6	60.0	59.0
9	33.4
10								
11	35.6	45.0	34.4	38.2	35.5	44.2	32.1	39.8
12	17.1	26.0	19.2	18.8	21.4	33.1	37.3	28.8
Medians	32.6	33.0	32.4	32.5	32.9	40.6	43.9	39.8

The results here presented might give the impression, at first, that the subjects of Group P progressed more rapidly after they come to the composition material. This they did for a short time, but the final marks show that this did not continue to the end. This supposition is all the more incorrect when it is known that cases one, three, nine and two started as late as the last few days of November. Because of this late start the results in these cases were so irregular or missing altogether that they do not do justice to the "composition method" to include them. Case four of his group could not be induced to use the touch method in her work. For this reason the results in this case are not included. Case, number ten, was dropped from the course on account of physical inability to acquire this skill.

The mortality in the work of typewriting for the two groups was decidedly greater for the cases in Group W; but it should be kept in mind that this group was made up of students who had already registered for a full course before taking up the typewriting. This made the work of this group very irregular.

CURVES OF ACQUISITION AND OF ERRORS, GROUP W

Case No. 11

Case No. 7

Case No. 3

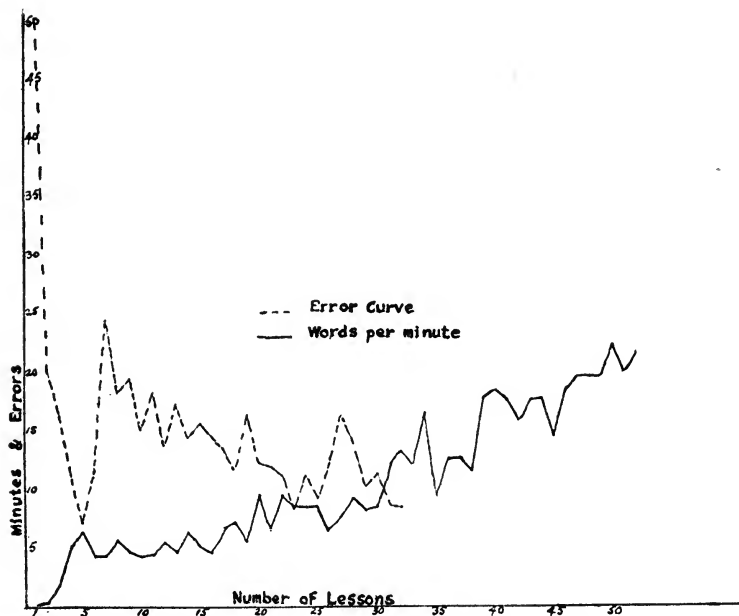
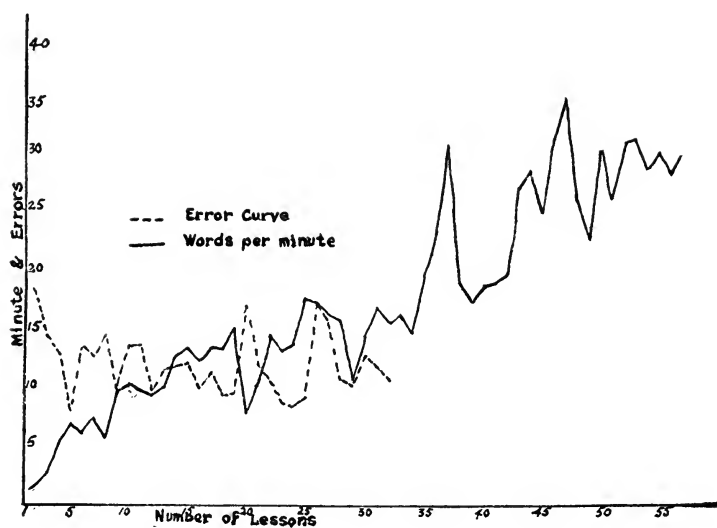
Case No. 2

Case No. 6

Case No. 5

It was impossible to obtain curves representative of the early learning of Group P that could, in any way, be compared with the ones here presented. This was due to the nature of the exercise material used, and since the two situations represent great differences in this matter it is needless to present such curves here.

The curves here presented are very much like those presented by other investigators of learning in this and in other fields, except that in the acquisition curves there is an absence of the initial rapid



rise, although there seems to be a rather abrupt upward direction about the fourth day. This is about the time most of them had mastered the keyboard sufficiently to begin doing copy work. By this time the student had worked himself out from under what, to him, had seemed an overwhelming difficulty.

Each student being taught by this "larger unit method" gave evidence of discouragement in her first few attempts at using the typewriter. This was hard to overcome. The complex of finding the proper key with the proper finger in the proper order by means of looking upon a chart, composing a sentence, keeping in mind the mechanics of composition construction, using back spacer, spacing properly, shifting carriage and attempting numerous other matters provides a situation which common sense has never been known to attack as a whole. This is what is asked of the student by this method of learning. It is likely that this overwhelming complexity is the factor responsible for the situation of "piece meal" in most of our work of teaching and learning.

This same practice of "piecemeal" is had in much of the work of secondary schools and in higher institutions. This is likely the situation for much of what is demanded as prerequisites. It might be found to be very economical to get the necessary mathematics for the later work of physics at the same time that the physics is given. It would be better to give the two together thus involving only the mathematics necessary to the other work. The same is likely true for the helps in English or any other subjects of curricula. Some such change as this is necessary before the arrangement of courses can ever conform to "beginning by doing a thing as nearly as possible in the way it is eventually to be done." It seems that at some future time the content of all work of formal education will conform much more closely to the actual life outside than it has ever done and that the method used in teaching will approach much more closely the ways of doing the thing after it is learned.

NOTES ON ARTICLES IN EDUCATIONAL PSYCHOLOGY IN CURRENT ISSUES OF OTHER MAGAZINES

REPORTED BY CECILE COLLOTON

Department of Educational Psychology, The Lincoln School of Teachers' College

INTELLIGENCE TESTS

Army Alpha and Student's Grades, Illustrating the Value of the Regression Equation. Homer Davis. School and Society, 1921, September, 223-227. Results of an experiment conducted at Stanford University to determine the value of the Alpha intelligence test as a means of predicting the type of work students would do; several case studies; advantages of the plan.

Criteria for the Regrading of Schools. James L. Stockton, Corinne Davis and M. Alice Cronin. The Elementary School Journal, 1921, September, 55-66. A program for efficient grading based on a central criterion of mental age supplemented by criteria of (a) chronological age; (b) physical age; (c) pedagogical age; (d) character age. Case studies and results of such a program in the Training School of San Jose State Normal School.

On the Need for Caution in Establishing Race Norms. Ada H. Arlitt. Journal Applied Psychology, 1921, June, 179-183. Report of an investigation to determine the relative influence of race and social status on the distribution of intelligence. Social status factor very important.

Estimating Intelligence by Means of Printed Photographs. L. Dewey Anderson. Journal Applied Psychology, 1921, June, 152-155. Results of an investigation to determine the reliability of photographs for indicating the intelligence of strangers. Correlation between assigned ratings and intelligence 27.

Two Cases Showing Marked Change in IQ. W. T. Root. Journal Applied Psychology, 1921, June, 156-158. Details of 2 Binet retests which showed distinct increase in IQ. Discussion of contributing factors.

The Predictive Value of Short Intelligence Tests. C. F. Hansen, M. J. Ream. Journal Applied Psychology, 1921, June, 184-186. The reliability and predictive value of short intelligence tests as determined by an experiment with two groups of students in the School of Life Insurance Salesmanship at Carnegie Institute of Technology.

Studies in Mental Tests. J. E. DeCamp. School and Society, 1921, October, 254-258. Results of testing Pennsylvania State College Freshmen with Army Alpha, Thurstone IV, and Stanford-Binet. Unreliable for predicting quality of collegiate work.

GIFTED CHILDREN

Gifted Pupils in the High School. John C. Almack and James L. Almack. School and Society, 1921, September, 223-228. Conclusions as to number of gifted children in high schools; the best means of discovering them; their physical

superiority; need for educational reorganization in their favor; and their superior social status; based on an investigation of the six upper grades of the Eugene, Oregon, schools.

Preliminary Report on a Gifted Juvenile Author. Lewis M. Terman and Jessie C. Fenton. *Journal Applied Psychology*, 1921, June, 162-178. The history of Betty Ford. IQ by the Stanford-Binet—188; by the Army Beta Test—175.

MEASUREMENT OF PERSONAL TRAITS

A Preliminary Study of the Correlations Between Estimates of Volitional Traits and the Results from the Downey "Will—Profile." G. M. Ruch. *Journal of Applied Psychology*, 1921, June, 159-162. Actual test scores of more than twenty graduate students on Downey "Individual Will—Temperament Test" compared with estimates for the same group secured from two groups of associates—university instructors and students in the same classes.

The Measurement of Aggressiveness. H. T. Moore, A. R. Gilliland. *Journal of Applied Psychology*, 1921, June, 97-118. Description of three tests for measuring aggressiveness; results of the use of the tests in an experiment at Dartmouth.

TRANSFER AND LEARNING

First Year Latin and Growth in English Vocabulary. W. L. Carr. *School and Society*, 1921, September, 192-198. Description of an experiment conducted in 7 schools to determine the effect of one year of Latin on a pupil's "passive" English vocabulary; results show Latin to be a definite aid.



What is the Disciplinary Value of the Classics? Thaddeus L. Bolton. *School and Society*, 1921, September, 205-210. An analysis of the term "mental discipline;" the classics as a tool with which to develop natural capacity; the possibility of using other tools for the same purpose.

The Feeble-minded Blind. Leila Holterhoff. *School and Society*, 1921, September, 174-179. An experiment in teaching the mentally defective blind; the need for special classes for such children in our schools and institutions.

TESTS FOR SPECIAL ABILITIES

A Study in Industrial Psychology. Tests for Special Abilities. Elsie O. Bregman. *Journal of Applied Psychology*, 1921, June, 127-151. Report of an investigation to develop tests for special ability—sales-clerks and clerical workers in a large department store.

NEW PUBLICATIONS IN EDUCATIONAL PSYCHOLOGY AND RELATED FIELDS OF EDUCATION



1. *An Advanced Text in the Psychology of Learning*.—In the opening sentence Pyle states that; “In this book I have tried to state everything that is known about learning.”¹ The writer has kept to his purpose and the product is a veritable encyclopedia, crammed with tables and graphs. If the reader expects to find it wearisome, he will be pleasantly surprised; Pyle has succeeded, as usual, in making heavy facts attractive. It is not however a book for lazy reading.

The author adopts at the start the Situation—Response hypothesis, and in many respects his system is similar to that of Thorndike; but he does not always succeed in being consistent with the underlying logic of his position, for example: a tendency to superimpose attention, attitudes, etc. as active forces upon the mechanics of learning.

In the crucial problem of the transfer of training, Pyle states essentially the view presented in Thorndike’s treatise, but opposes Thorndike on the equally crucial problem of the organization of the mind. In the former, the author points out that what are usually called attitudes, ideals, attention, methods of attack, etc., fall within the hypothesis of *identical elements*. In considering the theories of Spearman and Thorndike relative to the general intellectual factor, the author, while admitting the insufficiency of evidence, is “inclined to believe that there is a general learning factor and also a general intellectual factor, a factor operative in all intellectual processes.”

One chapter is devoted to a discussion of drill. After reviewing the evidence, the author states his position without equivocation: “The experiments leave no doubt of the great value of specific drill, of direct practice. There is no reason to beat about the bush, evade or come at it indirectly. I must know exactly what the skill is, have some good reason for desiring it, then I should practice it vigorously, regularly, directly.”

Completing the conventional topics, the three last chapters are devoted to fatigue, the relation of instinctive traits to learning, and illustrations of certain statistical procedures. The instructor who

¹ Pyle, William Henry: “The Psychology of Learning.” Baltimore: Warwick and York, 1921, pp. 308.

uses the book will appreciate the class exercises which follow each chapter and the extensive bibliographies. It is a very useful, complete and comprehensive book which will serve admirably as a text for a course following an introductory survey of Educational Psychology.

A. I. G.

2. *An Important Study of the Physical Growth of Children from Birth to Maturity*.—Every student of mental and physical development has appreciated the uncertainty of conclusions based on averages for different ages, and has deplored the lack of repeated measurements on the same individuals over a considerable period of time. This lack for physical growth Baldwin¹ has supplied in an important and almost monumental study. A similar study in mental growth is promised and it will be awaited with great interest.

The monograph gives, in Parts IV, V, and VI, an historical survey of 911 investigations in physical growth in this country and abroad, 643 comparative tables of measurements of infants, pre-school children, school children, and adults under thirty years of age, based on 5,385,400 recorded cases in various countries and a carefully annotated bibliography of 911 titles. This gives some notion of the heroic proportions of the study. It summarizes all that science knows or reasonably conjectures on how children grow physically.

Parts I, II, and III report Baldwin's own comprehensive data. Part I gives a complete description of instruments and technique in securing twenty-three standard measurements with illustrations of preliminary work done under the auspices of the Iowa Welfare Station in several cities of the state. Part II gives the mean growth in weight of white and colored boys and girls from birth to the close of the first year with numerous charts giving individual growth curves. The correlations between weight and birth and various periods up to the close of the year are positive but low, especially so at the end of the year period. Norms for height, weight, and weight-height index for the first year are set up, based on 9074 Iowa infants, and comparisons are made with French and German and other American data.

Norms for pre-school children in height, weight, and weight-height index are reported, based on 36,958 Iowa boys and girls between the ages of birth and six years. The results are from the extensive study by the Federal Children's Bureau in the Children's Campaign of 1918.

¹ Baldwin, Bird T.: *The Physical Growth of Children from Birth to Maturity*. University of Iowa *Studies in Child Welfare*, Vol. I, No. 1, pp. 1-411.

Perhaps the most interesting and valuable part of the study is found in Chapters V and VI which set forth the typical growth histories of children between six and seventeen years of age, illustrated by 400 individual growth curves in height, weight, breathing capacity, sitting height, chest girth, strength of right and left arms, and strength of upper back. Highly interesting are the individual synoptic profiles of growth in fifteen to twenty-two traits. Intercorrelations for the consecutive development of nineteen traits for the years from seven to sixteen and for height, weight, and breathing capacity on college girls for the years from seventeen to twenty have been worked out for the first time. In addition, the total correlations have been analyzed by the method of partial correlations. The coefficients of variability tend, on the whole, to decrease from seven to seventeen, and, in general, to be higher in boys than in girls. The chapters are a veritable mine of information and contain scores of important conclusions or generalizations which can not be detailed here.

Anatomical development, measured by radiographs of the wrist bones, was determined on sixty-seven boys and girls and correlated with height and weight. The correlations are very high. The anatomical development of disparate twins shows, contrary to the universal belief, very marked differences. Physiological age, as evidenced by the advent of pubescence or first menstruation, shows wide variations and low correlations with other traits.

With such an array of data it is evident that Baldwin's study will be the standard reference for some time to come, and that every student of psychology and education will want to possess it.

V. A. C. H.

3. *A New Reading Monograph*.—The field of children's interests in reading has been well covered in the recently published investigation of Arthur M. Jordan, Ph. D.¹ Previous studies are carefully reviewed in the introductory chapters.

Chapter II deals with the results obtained from the use of a questionnaire. Responses were obtained from 3,598 pupils in four cities. The tabulations show that the reading interests of boys and girls are far from identical and that in both cases some interests increase or decline with age. Every effort was made to secure uncensored state-

¹ Jordan, Arthur M.: *Children's Interests in Reading*. New York: Teachers College, Columbia University *Contributions to Education*, No. 107, 1921, pp.143.

ments and bona fide opinions of pupils and the conclusions drawn from the analysis of tabulations are set down in detail.

The investigator decided that more objective evidence could be obtained by observing the choices of children in public libraries and reports the results of extended observations in four libraries. The interrelation and correlation of these results with those obtained from questionnaires and the sale of books are discussed in a brief chapter.

L. Z.

4. *Two Noteworthy Psychological Contributions Resulting from War Work.*—(a) Among the psychological by-products of the war may be listed the recent volume on *The Scientific Measurement of Trade Proficiency*.¹ The extensive research financed by the government, under emergency conditions, facilitated classification of skilled personnel during the war. That the resulting tests may be a factor in industrial readjustment is the hope of the author. The volume shows the gradual development of technique in the construction and administration of trade tests: (1) the oral test; (2) the picture test; (3) the performance test; (4) the written test. The discussion is non-technical and readable. The principles underlying the selection of questions and the objective scoring of results are stated. The reasons given for the discarding of certain types of questions, make for a better understanding of the objectivity of such measuring devices. Many trades are covered and the carefully standardized tests submitted should not only function in the solution of personnel problems in industry but also demonstrate the practical value of a thoroughgoing scientific attack on the psychological phases of industrial problems.

(b) The work of the Psychology Committee of the National Research Council is also represented by the volume on army mental tests² published with the authorization of the War Department. The thoroughgoing nature of the work done by the co-operating psychologists in the construction and standardization of tests is further exemplified in the tabulations and graphical representations of Chapter II. Mental tests demonstrated their value in the selection of men for officers' training schools and other lines of service requiring special ability. They also facilitated the prompt recognition of men of

¹ Chapman, J. Crosby: "Trade Tests. *The Scientific Measurement of Trade Proficiency*." New York: Henry Holt and Company, 1921, pp. IX + 436. 1/4

² Yoakum, Clarence S. and Yerkes, Robert M.: "Army Mental Tests." New York: Henry Holt and Company, 1920, pp. XIII + 303.

extremely low intelligence. The military importance of segregating these men is obvious.

"The Examiner's Guide" used during the war is included as Chapter III and is followed by a report of tests given in Students' Army Training Corps with tabular comparisons of the results with those obtained in various educational institutions.

Chapter V deals with practical applications. Aside from the tremendous significance of "Mental Engineering During the War" the tests comprised in Chapter VI throw light on numerous educational and industrial situations. The services of the Committee were cut short by the signing of the armistice but not before demonstrating (1) the application of the principles of psychology to concrete military problems (2) the importance of co-operation in practical scientific service.

L. Z.

5. *Psychology Applied to Business*.¹—This is a book on the psychology of buying and selling with little of direct interest to the educational psychologist or teacher. It is written for the most part in a simple and popular style, describing the psychological factors that are of importance in influencing people to buy. There are numerous examples and applications taken from the realm of business, but there are also several others taken from the standard psychological literature, some of which seem a little removed from the main purpose of the book. In describing unconscious memory we meet again Coleridge's servant girl, which recalls the passage quoted by James and thereafter by many others. The author has made good use of what he calls the historical method in advertising, which is a measure of certain trends as illustrated by the change in percentages of different types of advertisements over a number of years. His use of this method is very ingenious, as, for example, when he obtains an indirect measure of truthfulness in advertising by counting the number of superlatives used ("best," "finest," etc.) over a number of years. The counting and charting of various items over a period of years might in like manner prove useful in measuring trends in educational procedure. Altogether it is an interesting and stimulating little book, and it is simple and direct enough to appeal to the circle of readers for which it was written.

R. PINTNER.

¹ Kitson, H. D.: "The Mind of the Buyer. A Psychology of Selling." Macmillan, 1921, pp. X + 211.

6. *Tests for Guidance in the High School.*¹—This is the first of a series of educational monographs planned by the Journal of Educational Research and edited by Dr. Buckingham, the editor of the Journal. He is to be congratulated upon the excellence of this first monograph, which has certainly set a high standard for future authors.

Dr. Proctor set himself the task of finding out to what extent mental tests might be useful for the guidance of high school pupils and he has dealt with the subject in a very sound and sensible manner, showing what he believes to be their value at the present time and making no inordinate claims for them. Tests alone will not solve all the problems of vocational advice and direction. They are to be used along with teachers' estimates of ability, records of school success, the vocational ambitions of the pupil, and, the author might very well have added, common sense. How such a combination may work is shown by the comparison of "guided" and "unguided" pupils, wherein we note the fewer number of failures in the guided group. Working from the army data as to intelligence and occupations, suggestions for vocational guidance, mainly of a negative kind are made. By following up the pupils through high school and on to college, the author has shown the selective influence at work. The median IQ for first-year high-school pupils is 105; for high-school graduates 111; for college entrants 116.

The symmetry of the monograph is somewhat marred by raising the question of the relation of a particular test to a particular subject, working this out for one test and one subject, and then leaving the matter hanging in the air. It is obvious that we must have correlations of each test with English and with other subjects before we can make any statement at all as to the value of the analogies test, which is the one the author has chosen to correlate. We are left wondering why the author suddenly stopped.

Of great value and interest are the mental age norms for the Army Alpha Scale. They differ at certain ages considerably from those given by the army workers. The army mental ages were derived a group of adults tested on the Stanford. The norms in this book are based upon "several thousand California school children," and should be more reliable than the army mental ages, which are absolutely conditioned by what adults can achieve on the Stanford Scale. It is a

¹ Proctor, W. M.: *Psychological Tests and Guidance of High School Pupils. J. of Ed. Research Monographs*, No. 1, June, 1921, Public School Publishing Company, pp. 70.

pity that an age distribution of the scores of all the cases tested was not included in the book.

This monograph is a valuable addition to the ever-growing library of books on tests, and should be in the hands of all those interested in mental examination and guidance of high school pupils. But why should the author append to this scientific piece of work a description of the tests written in the style of a publisher's advertisement, for example, "the only test yet published . . . ;" "ten well-selected tests," etc. etc.? The tests described do not need it.

R. PINTNER.

7. *The Influence of Work and External Conditions upon Mental and Motor Efficiency*.—In a monograph of 95 pages, Mr. Peaks¹ has summarized the important studies of the influence of time of year, time of day, weather, heat, humidity, etc., upon efficiency, considering the facts with regard to alleged long and short types of periodicity, together with a discussion of methodology and a presentation of original data. The author believes that the evidence favors a yearly rhythm marked by a depression mid-winter with maxima in Spring and Fall and a diurnal variation marked by an increase from morning to mid-afternoon usually a temporary depression at noon, but that weekly, twenty-three day, twenty-eight day and other alleged types of periodicity are not regularly found. Sufficient information was not available to enable the author to disentangle the multitude of possible causal factors. Fatigue, light, humidity, temperature, atmospheric pressure, meals, etc., are considered in turn, but none of these seem to seriously influence mental efficiency. The relative insusceptibility of the mechanisms involved to disturbance by these forces which do greatly affect our *feelings* of fitness make an interesting chapter in psychology which is written in convenient form in this monograph.

A. I. G.

8. *A Book for Teachers of Geography*.—The following approximation to a definition of the term "project" is given in a new book on the use of problems and projects in Geography:² "Projects consist in doing what pupils think it worth while to do. By means of them the subject of Geography is vitalized, because projects involve the active and

¹ Peaks, Archibald G.: Periodic Variations in Efficiency. *Educational Psychology Monographs*, No. 23, Baltimore: Warwick and York, 1921, pp. 95.

² Smith, E. Ehrlich: "Teaching Geography by Problems." Garden City: Doubleday, Page and Company, 1921, pp. XIX + 306.

motivated participation of the pupils in carrying them to successful conclusions. . . . It is the real need for objective illustration that makes the projects vital."

The purpose of the book as stated in the foreword is to help teachers "by assisting to vitalize the subject of geography." The author notes recent progress in the field and discusses the increasing significance of human geography in the solution of the problems of civilization. After a discussion of past and present school practice, he outlines progressive tendencies, lists necessary materials and discusses desirable procedures in the selection and solution of problems. Over one hundred pages are then given to illustrate problems. The appendix contains, among other things, valuable information concerning the accessibility of illustrative materials.

The author is evidently of the opinion that geography projects, as such, should have a definite place in the program. The following quotations show his position with reference to the reconstruction of the curriculum and the source of problems: "Teachers will do well to make an inventory of the work of the grade which the course of study demands, by laying out their term's problems and projects." . . . "Thoughtful and extensive reading forms the basis for teaching by problems and projects. Out of this experience teachers become prepared to construct problems." . . . "From a well selected bibliography, a teacher can construct a vast number of interesting problems."

The purpose of the book is manifestly ameliorative. It is addressed to classroom teachers, and does not go back to fundamental philosophic and scientific considerations upon which more thoroughgoing reconstruction must depend.

The author's purpose would be better served by the omission of irrelevant qualifying phrases and indirect implications. The sentence structure is often awkward and unless the reader stops to put the thought elements in more psychological sequence, the import of statements is lost. The following quotations are in point: "By 1880, for example, the factory had gained precedence over farming in England, even though at that time her farmers were making excellent wages, for the simple reason that means and methods of communication abroad so improved that she could obtain food supplies from the outside world cheaper than she could produce these at home." A book designed to interpret new movements and encourage teachers to adopt progressive methods, deserves more careful psychological editing.

L. Z.

THE JOURNAL OF EDUCATIONAL PSYCHOLOGY

Vol. XII

December, 1921

No. 9

IS THE RATING OF HUMAN CHARACTER PRACTICABLE?

(Continued from November)

HAROLD RUGG

The Lincoln School of Teachers College,

The conclusions from the study of official ratings and of the data collected at the Officers' Personnel School,¹ confirmed the conviction that another more elaborate and objectified experiment should be made in army camps. That was done in November 1918 at Camps Sheridan and Taylor. The experiment at Sheridan was organized largely for the purpose of working out the final procedure to be followed at Taylor. The data for both experiments are used.

One hundred and fifty-one officers cooperated in the study. They met in two groups, one of 93, the other of 58. Practically all of these officers were college trained men. Their average score on the Army Psychological Test was B⁺, indicating that we had a distinctly intelligent group of cooperating officers. I personally believe that the men engaged in the study represented as high a level of intelligence as will be found in typical groups of supervisory school officers.

How the "Experimental" Ratings Were Made.—Each officer constructed a rating scale in accordance with the following directions:

First.—A list of officers was compiled which contained 25 or more names, made up largely from officers who were present in the conference room. The greatest care was taken to see that the names of officers used on the scale were only those whom the rater felt qualified to rate. It was felt then, and proven later, that it was most important that each person used on a scale should be known intimately by the rating officer.

Second.—These names were arranged as nearly as possible in exact

¹ Presented in the November issue of this journal.

rank order of merit. This was done *independently* for each of the five groups of qualities on the scale. It is of the greatest importance that the scale be made separately for each group of qualities. From each of these lists five scale-men were selected. This was done by the very careful scrutiny and review of the two or three men at the head of the list on a given quality. From this the "15" men were selected. Similarly, the "3" men at the foot of the list were selected from the two or three poorest; likewise, with the others, the "12," "9" and "6" men.

Third.—There was a thorough discussion, taken part in by all present, of the construction of the scale and, between the first and second conferences each rating officer studied and revised his scale.

It is my judgment that the Camp Taylor scales represent excellence in scale-making that it will be difficult, if not impossible, to duplicate under working conditions in public schools. It is important to remember this in applying the conclusions of the army investigation of rating in school practice.

Fourth.—Each of the 151 officers gave me a copy of his scale in order that a detailed comparison might be made of the extent to which officers agreed who had used the same scale-men. Remember, that the officers had constructed their scale, so far as possible, from their associates in the conference room.

Fifth.—At the second half-day conference each officer rated on his new scale each other officer in the room whom he was positive that he knew well enough to rate. No officer was allowed to rate another unless he was sure that he was thoroughly qualified to do so. To make this effective and to supply statistical information, each officer stated the number of weeks that he had associated in the army with the officer rated and estimated the extent to which he was qualified to rate him. In addition, each officer submitted a list of the other officers present, who, in his judgment, were competent to rate him. These officers are referred to in my subsequent discussions as "competent raters."

Sixth.—A third half-day conference was held at which this entire procedure of revising scales and ratings was repeated. Thus, two conference ratings are available on each officer which might be properly regarded as made by officers who were thoroughly competent to rate.

Seventh.—Each officer supplied detailed personal data concerning his age, schooling, college honors, annual earnings, occupational activities, etc.

Eighth.—Each of the officers took the Army Psychological Test once, a number of them twice, and all of them took the Thorndike Alertness Test twice.

Ninth.—A round-table discussion was held in which the officers stated the difficulties which they had encountered in constructing and using scales and in which they made recommendations for changes and improvements in the procedure.

HOW CLOSELY DO RATINGS AGREE WHEN MADE BY DIFFERENT RATERS ON THE SAME PERSON?

From this elaborate experiment what does one learn about the reliability and practicability of rating human character? *He learns that even under such carefully controlled conditions, it is practically impossible to secure ratings on point scales which are reliable estimates of character.* He learns first, for example, that when a person is rated independently by from three to thirteen competent raters, that the range in the ratings will commonly be as large as 30 points on a total scale of 80 points. In but one case in a list of 16 officers was the extreme difference less than 20 points. In the case of 8 of the 16 the range exceeded 30 points. Table II summarizes the data.

TABLE II.—AVERAGE OF THE RANGE AND AVERAGE OF THE AVERAGE DEVIATION OF 3 TO 13 INDEPENDENT RATINGS ON A PERSON

	Average of the range				Average of the average deviation			
	General conference group		Raters reported as competent to rate by those rated		General conference group		Raters reported as competent to rate	
	Very well qualified to rate	Less well qualified	Very well qualified	Less well qualified	Very well qualified	Less well qualified	Very well qualified	Less well qualified
A. Includes all persons rated by 3 or more raters.....	16.9	17.6	15.6	21.5	6.3	7.7	7.0	8.4
No. of cases.....	28	23	21	6	28	24	22	7
B. Includes only persons who were rated by 3 or more raters in both groups.....	18.7	19.9	10.7	21.7	6.7	8.9	4.5	4.2
No. of cases.....	18	18	3	3	19	19	2	4

One learns furthermore that the typical variations in such independent ratings (as shown by the average deviation) are commonly in the neighborhood of 7 points (for the data of the first conference it was 7 points and at the second it was 6). The probability is too uncertain, therefore, that a single rating on a person will locate that person even within his proper "fifth" of the rating scale. *The chances can not be more than 4 to 1 that any rating will be within 14 points of the person's true rating.* These results, therefore, confirm strikingly, the results obtained from the Fort Sheridan ratings and from those secured in the first experiment at Camp Sheridan.

With such conclusions before us two questions press insistently for answer. First: Why do independent ratings made so carefully disagree so widely? Second: Is the process of judging human character so nearly impossible of objectification that "agreement or disagreement" between ratings is not an adequate criterion for measuring the reliability of a scale? If two raters assign the same score to a person do these scores represent equivalent and comparable judgments of character? We have the data at hand upon which to base a rather careful discussion of these questions.

WHY DO INDEPENDENT RATINGS VARY SO GREATLY?

Four possible causes of disagreements were explored:

1. Lack of acquaintance.
2. A tendency to rate high or low.
3. The extent to which the analysis of "military abilities" (which is involved in the construction and use of rating scales) rests upon mental backgrounds that are so distinctly different as to result in totally different placements of the same officer on different scales. It was recognized that such an analysis may even result in similar placements, and quite different evaluations of ability. Thus, two independent ratings on an officer, as well as two different scales, may represent a very different distribution and differentiation of ability.
4. Is the process of discriminating clearly the elements of human character contributed to by so many complicating factors as to render practically impossible close agreements in total estimates of the man? These complicating factors may include: (a) the ability to state objectively outstanding personal qualities; (b) the effect of peculiarities of special groups of qualities on the estimate of other qualities;

(c) the emphasis given to one group of qualities by one officer and to a different group by another; (d) the complexity of the task of holding in mind the various special traits which contribute to a group of qualities without elaborate devices for discriminating the subordinate elements.

I shall treat the third and fourth factors in great detail. The discussion, however, of the "effect of extent of acquaintances" and of a "tendency to rate high or low" must be inadequate. Data for the former are incomplete because of our care to permit only competent raters to rate. However, there were differences observable in "extent of acquaintance" and I report the results as suggestive of more sweeping conclusions that I believe would obtain from more complete data.

1. *The Effect of Acquaintance.*—Comparison was made of the agreements of ratings for two groups of raters representing different degrees of acquaintance. The difference between the average deviations of the two "acquaintance" groups is rather marked. The first set gave 8.9 for one group and 6.7 for the other; the second gave 7.7 against 6.3. Furthermore, the average deviation for the "competent raters" was smaller than for any other group. This we found to be 4 points. The data that we studied during the investigation were impressive pointing toward the conclusion that estimates of human character depend closely upon intimacy of acquaintance and that it is important to evaluate the competency of the rater. Another investigation has been reported to me in which the reliability of letters of recommendation was studied. The accuracy of the recommendation was correlated with certain objective methods, and the correlation ranged with the intelligence of the "judge" from 0.7 to 0. Evidence like this points to the importance of both intimacy of acquaintance and competency of judgment.

2. *The Tendency to Rate High or Low.*—The dearth of ratings by any single officer caused great difficulty in treating this factor thoroughly. If the average rating had been based upon a reasonably large group, say 25 or more, a correction could have been applied by the following procedure: (a) the determination of the variability (by σ or by Q) of each rater's ratings; (b) the average of the ratings falls at 60; the estimated standard deviation is between 13 and 15, and the estimated Q is 10; (c) the ratings of each rater thus could be approximated by equating the position of each of his actual ratings on the scale to the position that it would have occupied on the scale

of the "true" distribution. The average would be made 60 and each other rating would be increased or decreased by the relative difference between it and the average.

To investigate this factor thoroughly the ratings of persons should be available, each person rating a considerable number of individuals. Only four officers in my study rated as many as 10 to 14 persons, and only one rated 15. From the experience gained in carrying through this inquiry it is doubted if, either in the army or in school practice, an experiment can be set up in which a larger number of valid ratings can be obtained on one person under natural rating conditions. For these reasons the analysis of the tendency to rate high or low included merely a careful approximate correction of ratings made by extremely high and extremely low raters.

The results were all plotted and submitted in the original report. Lack of space prohibits their publication here. I shall ask the reader to accept my conclusion from the careful study of these meager data that divergencies in judgments of character are but slightly, if at all, accounted for by a tendency to rate high or to rate low. Hence, the great importance of studying carefully the third and fourth factors referred to previously.

3. *Incomparable Scales as Contributing Causes of Differences in Rating.*—Of far greater importance is the question: Are the scales comparable on which ratings are made? Do the mental backgrounds against which estimates of character are made afford similar standards of judgment. Several basic questions must be answered:

(a) When an officer is used by two or more raters, is he assigned the same scale value on the different scales? Is he a "fifteen" man on all scales, a "twelve" man, a "six" man, etc.?

(b) Does the placement of scale-men on the "intelligence" part of the rating scale correspond closely to placement in intelligence by the Army Psychological Tests?

(c) Are differences in two ratings on an officer by different raters paralleled by corresponding differences in the positions of "scale-men" on the scales used in the rating?

(d) Is the same officer used as a "scale-man" on more than one group of qualities?

ANALYSIS OF THE "SCALE-MEN" ON 45 RATING SCALES

To answer these questions I made a minute analysis of 45 rating scales. The first step was to determine the extent to which a given

scale-man was used at the same scale-value by different persons. Fortunately a large number of cases are available. Four hundred and fifty-eight instances were found on the scales in which the same men were used on from 2 to 13 different scales. Each instance of recurrence of a name on different scales was tabulated to give a semi-graphic chart upon which the following facts were shown:

1. The specific scale-value at which the officer was used on each scale.
2. The Psychological Test score and the Alertness Test score of each officer making the scale.
3. The Psychological Test score of the officers used on the scale.
4. The total number of scales upon which the officer was used at each scale value of each quality.
5. The per cent of the total number of times an officer was used on a given quality on which the consensus of agreement was used.
6. The average position to which the officer was assigned on each quality.

Conclusion.—It is impossible to print in this article the ten tables or the detailed charts which were made up to study this question. In Table III a summary is presented of the essential data. A minute analysis of the original tables and charts shows that rating scales made under as carefully controlled conditions as were those, are distinctly incomparable. *Intervals upon different men's scales do not represent closely the same amount of the trait in question.* Table III (typical of three other tables) shows that *in not more than 50 to 60 per cent of the cases will an officer appear at the same scale value on different scales.* It shows further that a *divergence of $4\frac{1}{2}$ to 6 points (that is, 35 to 50 per cent of the entire scale!)* will typically occur between different placements of the same officer. It shows that the same man, used at "15" on one scale, will be used at "12" in another, "9" on others, "6" on occasional ones and even "3" on a few. Persons regarded as the "best captain I ever knew" were selected for scale positions on other scales as "the poorest captain I ever knew!" And the selection was made by the most objectified procedure we have yet devised and by a procedure that we probably cannot duplicate under practical rating conditions in education.

But, in using the "agreement" of scale placement as our criterion, we are employing subjective methods. It happens that we have at hand a very good comparison of subjective "ratings" with objective measures of a particular group of traits,—namely, "intelligence" as

measured by the Army Psychological Test and by the Thorndike Alertness Test. The comparison of ratings and actual performance scores of 59 officers is presented and summarized in Table IV. These

TABLE III.—PER CENT OF PERFECT AGREEMENT IN SCALE-POSITIONS, TOGETHER WITH AVERAGE NUMBER OF UNITS OF DIVERGENCE FROM THE "CONSENSUS" OF AGREEMENT

(All names used occurred on 5 or more Scales)

Each row represents data on one officer

	Physical qualities		Intelligence		Leadership		Personal qualities		General value	
	Per cent	Average deviation	Per cent	Average deviation	Per cent	Average deviation	Per cent	Average deviation	Per cent	Average deviation
	100	...	78	6.0	50	7	57	8.0
	75	3.0	50	4.0	100	57	14.4
									71	12.0
	83	3.0	50	3.0						
	80	3.0	50	3.0	50	5.0	50	5		
	60	3.0								
	50	6.0	60	6.0				
			60	4.5	50	4.0				
			50	8.0						
	100									
	40	8.0	40	7.0	60	24.0
									80	8.0
			60	3.0						
			40	3.0						
	86	3.0			80	6.0	67	8.0
					20	4.5			60	8.0
					40	4.0				
					60	3.0				
	60	6.0								
			80	9.0	50	13.3
	60	4.5								
Average....	72	4.4	56	5.1	58	4.6	50	6	63	11.9

two sets of measures were made comparable by equating the range and variability of the "ratings" to the range and variability of the test scores. Thus, a "fifth" of the test "scale" is comparable to a "fifth" of the rating "scale." The detailed tables of the original report gave the following facts: (1) the average position on the intelligence part of the rating scale for men of given test score standings, say "15," "12," "9," "6," and "3;" (2) the converse of these data.

We can now answer the question: Is "intelligence" discriminated accurately on the man-to-man comparison scale (assuming the validity of our test measures of "intelligence")?

It appears from Table IV that the men selected as "most intelligent" ("15" men) were rather accurately picked. Of 13 men, 4 are rated "15," 8 are rated "12," and one is rated "9." For the other four scale positions the discrimination is very inadequate. Men of "average" intelligence ("9") as shown by the tests are rated higher than "superior" persons ("12"). The average test scores for the groups denoted "12," "9," "6" and "3" on the rating scales, are respectively 9.24, 10.29, 9.3 and 8.46.

We find at the "lowest" end of the scale great differences between the rating on "intelligence" and the test score on intelligence. Out of 28 "3" men on the rating scale only 5 appear as "3" men on the psychological tests; 5 appear as "6" men; 11 appear as "9" men on the test; 4 are placed at "12" on the test and three are "15" men on the test! No evidence which has been secured in this investigation makes more apparent the fact that the rating scale is thoroughly *relative*. "The least intelligent man I ever knew" according to the rating scale varies all the way from "the most intelligent" to "the least intelligent" as shown by psychological tests. The fact that psychological tests demand abilities which are not considered in the rating scale is not forgotten in this connection. Admittedly it should cause some lack of agreement between the test placing of a man and the placement by the rating scale. The two measures, however, clearly involve a very considerable number of common abilities. It is these common abilities that should operate to give a fairly close agreement between the two scales. The psychological test at least classifies together, reasonably well, officers who are nearly equal in ability. The rating scale should do this also. Hence, whether we expect agreement in scale placement or not, we would demand the same relative degree of homogeneity to come from the use of either scale.

With fewer cases, and hence with not so much reliability, the Camp Sheridan material confirms the judgment as to the "relative" aspect of the rating scale. It is this relative aspect of the scale that makes rating so different from test scores. It is believed that such an analysis as we are making, supports our earlier hypothesis that *the scales themselves do not represent equivalent amounts of the trait*. In this case, also, we find that the chances are not large that any one assign-

TABLE IV.—COMPARISON OF THE PSYCHOLOGICAL TEST STANDING OF OFFICERS AT CAMP TAYLOR WITH THEIR RATING SCALE-STANDINGS

(Standings in the two scales are expressed in "fifths of the respective scales."
Numbers in the compartments of the table give the psychological test fifths)

Number of officer making the scale	Psychological test of officer making scale	Alertness test of officer making scale	Rating scale fifths				
			(If test scores agree with ratings the numbers in these columns will agree with the numbers at the head of columns)				
			1 "Lowest"	2 "Low"	3 "Middle"	4 "High"	5 "Highest"
1	5	5		
2	4	..	5	3			
3	4	..	4	4		5	4
4	3	..	2				
5	4	3	5	4	3		
6	4	..	4	4			
7	3	4	5	3	
8	3	4	4	4	4		
9	2	..	3	3	1	4	
10	..	3					
11	5	..	3	..	4	5	
12	..	4	4	..	5	1	
13	..	4	5	3	
14	3	4	4		
15	5	1	3		
16	4	3	4	2	
17	5	..	3	2	4	3	
18	3	..	3	..	2	..	4
19	5	4	3	..	4	4	4
20	4	5					
21	1	..	3	1	
22	3	2	2	1	5
23	1	1	4		
24	4	..	3	..	4	..	5
25	..	3	1	3	2		
26	..	4	1	5	4	4	
27	4	5	3	5	4
28	3	2	2		
29	5	4	..	3	4	2	4
30	..	3	3	..	4		
31	..	1	4	2	3
32	2	4		
33	4	..	3	1	3	5	
34	5	..	3	..	2	..	5
35	..	2	5		
36	2	2	4	4
37	4	4	..	2	2		
38	4	2	
38A	2	3	
38B	3		
39	5	4			
40	4	5	..				
41	..	5	3	
42	3	2					
43	4						
44	3	1	1	..	5	2	4
45	5	4	4	
48	1	..	1	..	2	4	
49	..	3	4		
50	..	4	4
51	1	..	1				
52	4	2		
54	5	..	2	2	5
55	4	..	2	..	4		
56	1	..	3				
57	2	1	
Average..	2.82	3.1	3.43	3.08	4.23

ment of an officer to a definite scale position will represent closely the true scale placement of the man.

In the course of our tabulations we studied the question: "Do raters of superior intelligence discriminate *intelligence* as mental tests do?" The answer is: "Slightly better than raters in general but yet not at all well." Tables were made comparing ratings and tests score of the men used on scales by officers of superior intelligence. Of 20 "5" men, 8 (40 per cent) were assigned to exactly the same position on the rating scale as they obtained in the test. But the average displacement for the whole group was 1.25 intervals, or 3.75 units. Thus, taken as a group, officers of outstandingly superior intelligence discriminate intelligence in others little better than do those of "average" intelligence. The evidence we are piling up points constantly to the very great difficulty of estimating character. It certainly calls for further and more minute analysis of the process of judging human traits.

Two points are clear then: first, most careful construction of the man-to-man comparison rating scales does not lead to close agreement in placing scale-men; second, experimental "ratings" of ability show distinctly inadequate agreement with test measures of ability.

If a number of persons disagree widely in rating another person, it may be due, wholly or in part, to differences in evaluating the abilities of persons used at particular positions on the rating scale. We are ready, therefore, to study in detail the third important matter.

ARE DIFFERENCES IN RATINGS PARALLELED BY CORRESPONDING DIFFERENCES IN THE PLACEMENT OF SCALE-MEN?

In the course of the discussion of this question we shall answer two related ones of the greatest importance: (a) if two ratings on a person agree closely, can it be inferred that each total rating is contributed to by approximately the same estimate of the component traits; (b) if scales can be found which do, throughout the range, represent nearly equivalent amounts of any trait, will ratings upon these scales closely approximate each other?

These questions are of the highest psychological importance. Very elaborate plates and tables were prepared to answer them. This section of the original report, as submitted to the Army Committee on Classification of Personnel, discussed the situation for a group of 10 officers. These 10 were selected for the comparison solely because they

were rated by eight or more different raters. They provide an important opportunity to study correspondence in overlapping scales. (I shall use "overlapping scales" to mean "scales which used one or more common names.") It is certain that these ten are representative of the entire group.

TABLE V.—PER CENT OF 45 CASES IN WHICH THE SAME PERSON WAS USED ON 2 SCALES FOR WHICH HE WAS ASSIGNED THE SAME VALUE

29	60
58	50
36	21
36	52
37	37

I can only print summary data, Tables V and VI. (The details are available to any student of the problem.) Table V shows that in only one-third (37 per cent) of the cases in which an officer is used on a scale will be assigned the same value. Table VI presents a summary of the

TABLE VI.—COMPARISON OF THE DIFFERENCES IN TOTAL RATING OF TWO RATERS ON ONE OFFICER WITH THE DIFFERENCES IN THE VALUES TO WHICH THEY ASSIGNED COMMON-SCALE-MEN

Officer's number	Second rating showed increase of 0-5 points over first rating				Second rating showed increase of 6-10 points over first rating			
	Number of instances in which the man was assigned to		Total numerical difference between the scale values on the two scales on which same men were used at differ- ent values		Number of instances in which the man was assigned to		Total numerical difference between the scale values on the two scales on which same men were used at differ- ent values	
	Same value on 2 scales	Different values on 2 scales	First rating	Second rating	Same value on 2 scales	Different values on 2 scales	First rating	Second rating
1	20	7	68	70	6	5	51	36
9	12	9	91	88	10	9	129	94
12	3	2	20	22	2	1	16	8
13	4	6	106	87	1	12	138	162
17	1	13	191	231	4	1	6	15
18	5	5	71	55	1	5	74	106
21	3	2	9	12	6	5	57	45
27	7	11	147	174	10	11	127	148
28	1	4	7	102	102
29	4	17	176	197	6	11	92	98
Total.....	60	72	50	67		
	132				117			
	45.4 %	54.6 %	42.7 %	57.3 %		

TABLE VI (Continued).—COMPARISON OF THE DIFFERENCES IN TOTAL RATING OF TWO RATERS ON ONE OFFICER WITH THE DIFFERENCES IN THE VALUES TO WHICH THEY ASSIGNED COMMON-SCALE-MEN

Officer's	Second rating showed increase of 11-20 points over first rating				Second rating showed increase of more than 20 points over first rating			
	Number of instances in which the man assigned to		Total numerical difference between the scale values on the two scales on which same men were used at different values		Number of instances in which the man was assigned to		Total numerical difference between the scale values on the two scales on which same men were used at different values	
	Same value on 2 scales	A different value on 2 scales	First rating	Second rating	Same value on 2 scales	A different value on 2 scales	First rating	Second rating
1	8	9	187	143				
9	14	6	57	63				
12	5	3	12	33	1	2	15	27
13	3	17	213	224				
17	8	14	156	156				
18	1	2	25	14				
21	22	19	205	165	11	13	160	149
27	7	12	113	174	4	10	101	128
28	1	3	45	61	6	10	102	148
29	1	2	9	18	2	2	18	27
Total	70	97	24	38	Total number of overlappings = 478	
	167		62			
	41.8%	58.2%	38.7%	61.3%		

data on 478 instances of the use of the same person on two scales. The detailed exhibit and Table XVIII lead clearly to the conclusion that differences in total rating are not parallel by even approximately corresponding differences in scale-values.

Differences in assignment of scale values are discernible only when the differences in total rating become very large. This leads me to suggest an explanation for differences in rating: that the whole operation is contributed to by many small differences in estimating the presence of subordinate groups of qualities; that many of these small differences neutralize each other in contributing to the total score, some being increases and some decreases; but that when the differences in estimate become very large, it can be shown statistically that these differences in placing men on the scale tend to parallel differences in rating on the scale.

It is clear, therefore, that we need to explore in great detail the comparison of individuals with scale-men. In presenting the individual cases we shall answer the question:

When two persons agree closely in the total rating to be assigned another, can this agreement in numerical rating be interpreted to mean an agreement in judgment of character? I believe this inference can be drawn only when the total rating is contributed to by equivalent estimates of component traits against equivalent scale-men. Specifically, if two persons give another person the same rating, say 72 and 72, are the ratings contributed to by like agreements in comparing the person rated with the respective scale-men? Only by such coincidence in rating subordinate traits against equivalent standards of judgment (the scale-men) can we infer that agreement in total rating represents equivalence in estimates of total character, and not a mere chance situation.

We need some specific cases to get the matter before us clearly. I shall discuss two distinctive conditions: the first, that in which two scales agree, that is, that the two scales consist of identical names at several values; the second, that in which the scales are apparently very different (the same person appearing on the two scales at different values), but in which ratings made against them are the same. The scales of 59 persons have been compared, value by value, and all instances noted in which a given person appears in two or more scales. (Exhibits VIII and IX of the original report give the scales themselves.)

It was very difficult to set up natural conditions for constructing scales and for rating persons against them which at the same time would insure many instances in which, on two or more scales, the same scale position would be assigned to a given individual. We are fortunate, in fact, *to have one instance of two scales, those of officers No. 11 and No. 24 in which the same officer appears at the same value in 8 of the 25 places on the scale. Furthermore, No. 11 and No. 24 rated 9 persons in common.* Hereafter I shall refer to particular rating officers by number. Here we have, for 9 persons, two independent ratings made against standards of judgment which presumably are based on like judgments of character. Is it not of great importance that *this is the only instance (in 59 cases) in which I have been able to show that scale construction is based upon equivalent judgments of character?*

The scales of No. 11 and No. 24 for "physical qualities" and for "intelligence" are reproduced exactly in Table VII. Note that the names or numbers of each scale-man and the numbers of each officer rated on both scales, are located at the proper value (from 15 to 3). In "physical qualities" the "15," "9" and "3" men are the same on both scales.

TABLE VII.—COMPARISON OF "PHYSICAL" AND "INTELLIGENCE" SCALES CONSTRUCTED BY NO. 11 AND NO. 24 TOGETHER WITH THEIR RATINGS ON SAME OFFICERS

Psycho- logical test stand- ing in scale fifths	Average position on others scales	Average of con- ference ratings on him	Scale values	No. 11's scale numbers of scale officers	Ratings assigned by No. 11 to offi- cers rated by both No. 11 and and No. 24	Scale values	No. 24's scale numbers of scale officers	Ratings assigned by No. 24 to offi- cers rated by both No. 11 and No. 24
The Physical Qualities Scale								
1	7.8	63.0	15	Whitcomb	No. 36	15	Whitcomb	No. 36
			12	No. 23	No. 19, 17, 12, 8, 37	12	Swift	No. 19, 17, 8 No. 37
			9	Whiting	No. 22	9	Whiting	No. 12
2	8.3	63.0	6	No. 32	Staker	6	Rundle	No. 22
			3	Shippen	No. 21	3	Shippen	Staker No. 21
The Intelligence Scale								
*5	10.5	64.5	15	Swift	No. 36	15	No. 11	No. 36
5	10.8 12.0	82.0 63.9	12	No. 17	No. 19, 17, 37, Staker	12	No. 1	No. 19, 17, Staker
								No. 8 No. 12
3	9.0	67.0	9	No. 8	No. 12, 8	9	Goodnow	No. 22, 37
*	11.0 11.0	55.6 55.6	6	No. 7	No. 22, 21	6	No. 7	No. 21
3	4.3	51.8	3	No. 4		3	No. 4	
*3	4.3	51.8						

* No. 11, 1, 7, and 4 on No. 24's scale.

On the "intelligence" scale the "6" and "3" men are identical. On the Leadership scale the "9" man is the same on the two scales. On Personal Qualities the "15" man is the same. And, finally, for "General Value" the "16" man is the same on the two scales. These eight names are common to the two scales and are assigned to the same scale-value. Here we have a case of two scales in which the Physical and Intelligence qualities may be regarded as representing about the same "spread" or differentiation of the traits in question and in which

a unit interval on one scale must represent closely the same difference in amount of the trait as a unit interval on the other.

Table VII reproduces the ratings on each quality and the total ratings given each of 9 officers on these two scales. Note that the differences in total rating are, respectively, 4, 3, 14, 4, 4, 4, 1, 9, 2, and the average difference for the 9 ratings is 5 points. The large divergence of 14 in one case is caused by a difference in rating on General Value in the case of No. 21, No. 11 having rated him 8 on General Value while No. 24, rated him 23. The other large disagreement in total rating, that of 9 points, occurs on Personal Qualities, the largest difference in a single quality being caused by one officer being rated 12 by No. 11 and 7 by No. 24. *The physical scale is the only one that we are reasonably confident represents equivalent distribution of the trait and it is on this that there is almost perfect agreement in rating the 9 men.* The differences between the physical rating given these 9 officers by Nos. 11 and 24 are, respectively, 2, 1, 0, 1, 2, 1, 1, 0, 1. Furthermore, they are rating men who are distributed throughout all portions of the total scale. *It is significant that the two ratings on a single man agree, irrespective of the amount of the trait that he represents.*

The same almost perfect agreement in rating men on these scales is found when we study the intelligence qualities. In only one of the 9 cases is the divergence in rating intelligence more than 2 points. (The case of No. 37, a Major, rated on a Captain's scale. It is worth noting that in rating Staker, another Major, the agreement on him by the two raters is within 1 point in 4 of the 5 qualities and within 2 points on General Value.)

The study of such comparable scales shows that *it is possible, but extraordinarily difficult*, for two rating officers to construct scales, the intervals of which will represent approximately equal differences and to make the man-to-man comparison that is necessary in supplying a total judgment of a person's worth. In this analysis, *the case that we have just discussed is the only one which has been found in which the construction of scales and ratings made upon them leads to a satisfactorily close agreement in the estimation of character.* This instance, by being an exception to the rule, throws into sharp relief the fact that the estimation of character involves very striking difficulties. In our consideration of further examples, however, we should hold in mind the fact that two officers have constructed scales and estimated the presence or absence of complex traits on 9 other men with very close agreement. We should also remember that equal total ratings can be

interpreted to be equivalent estimates of character only when the two scales against which ratings are made, represent closely identical scale-values and "spread" of character.

(In the January issue further detailed examples will be given of scale construction and of rating, together with a more extended interpretation of the data.)

RATE OF MENTAL GROWTH, AGES NINE TO FIFTEEN

FOWLER D. BROOKS

Johns Hopkins University

The results of annual re-tests of one hundred seventy-one children, ages nine to fifteen, of various economic and social groups, (as to mental ability, more highly selected at the earlier ages than at the later ones) give evidence that the rate of mental growth, as measured annually, is very nearly a straight-line affair, and is approximately the same for each year for school population of these ages.

In May, 1918, 1919, and 1920, at the Training School of the Mankato, Minnesota, State Teachers College, a battery of the following tests¹ were given to grades IV to IX, inclusive:

Four Woodworth-Wells number-checking tests; quality of handwriting on two specimens of written work handed in by the pupils in other subjects such as language, history, etc.; quality of handwriting and speed of handwriting from two tests given a week to ten days apart; sixty words from columns Q, S, and U of the Ayres scale, dictated slowly in easy sentences, during three or four spelling periods on as many days; Thorndike Reading A2 and B, the x and y lists being given in alternate years, and only comparable parts of tests being used so that perfect score would be the same each year; Courtis Arithmetic, Form B, four fundamental operations, attempts and rights; Woody Arithmetic, Series A, four fundamental operations; Stone Reasoning; Composition; Woodworth-Wells Opposites; Pintner-Toops Revised Directions Test; Immediate Auditory Memory, Concrete and Abstract, Whipple's lists; Memory for English equivalents of Italian words—nine different tests devised by the writer and three given each year; Woodworth-Wells Substitution—five geometrical forms; Letter-Digit Substitution, devised by the writer and the same three tests given each year; a reasoning test, part of an Omnibus test devised by Thorndike and McCall; Trabue's Language Completion, scales C, B, and D—one each year; Thorndike Reading, Alpha 2. In 1920, Army Alpha and Thorndike Group Intelligence Test III, Series L, were given.

¹A complete account of this investigation is contained in a recent publication of the Bureau of Publications, Teachers College, N. Y., entitled "Changes in Mental Traits with Age, Determined by annual Re-Tests."

One hundred four of the subjects were given two annual tests; sixty-seven were tested annually three times.

Treatment of Data.—Each child's score in each test was subtracted from his score in the same test the following year. This difference represents his improvement in one year and is positive or negative. For those tested three years the same procedure was followed, taking the difference between the first and second, and second and third testings. These improvements in each test have been grouped according to the ages of the children. May 15th, was the median date of testing each year, and so ages have been computed as of that date. The age given in all cases, is the child's age on his last birthday.

In each test the median yearly improvement in gross score for each age and for each sex has been computed. On account of the unreliability due to the small number of cases and to some of the tests themselves, no comparisons have been made between the results of the individual or separate tests. We can secure greater reliability by combining the test-results into three or four groups of similar functions. A difficulty in combining or comparing results of different tests is the different-sized units. For example, is an improvement of 18.4 problems in Woody Arithmetic more or less than an improvement of 6.2 problems in Curtis Form B? Norsworthy, Thorndike, Woodworth, Sleight, and others have used the procedure of expressing gross score points in different tests as functions of some measure of the variability of the respective tests. This tends to equalize the units. We have divided the median gains in gross score in each test by the arithmetic mean of the standard deviations of eleven, twelve, and thirteen year olds for each sex in that test.

These median sigma gains in the separate tests could now be compared, but it is doubtful if any conclusions so drawn would be of much value. For example, in the handwriting in ordinary written work the boys made much greater gains from thirteen to fourteen, than from twelve to thirteen, while in the handwriting tests their improvement was at a constant rate for these two years. One could attempt to explain this by saying that the improvement in handwriting as measured by the tests became so permanent by thirteen that it carried over into ordinary written work to a greater extent than before this. It seems wise, however, not to lose sight of more probable causes—the unreliability of single test results. Accordingly, we have combined the median sigma gains in the different tests into four groups of similar functions, by finding their arithmetic means. The four

groups of functions we may designate *simpler*, *memory*, *higher* or *more complex*, and *informational functions*. We have also found the arithmetic means of the median sigma gains of all the tests, giving us a composite of more than forty different tests. These are given in Table I.

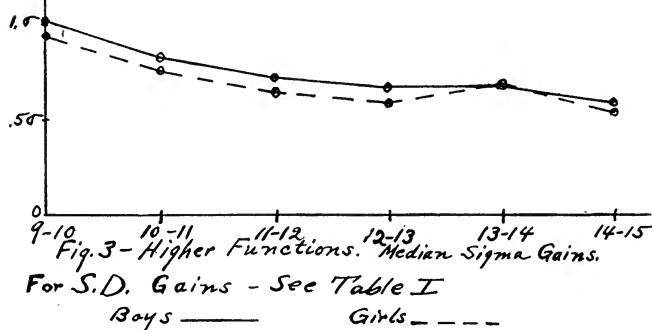
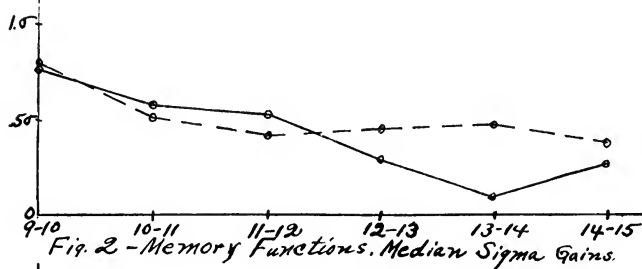
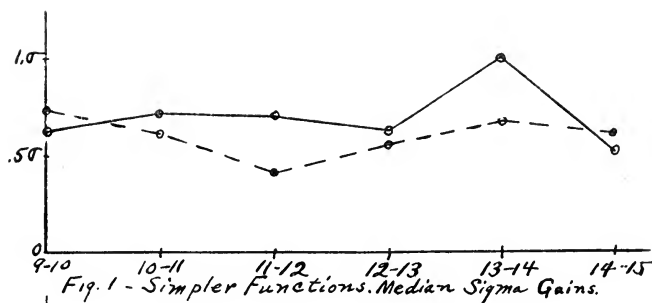
TABLE I.—MEAN GAINS IN SIMPLER, MEMORY, HIGHER, INFORMATIONAL, AND COMBINED FUNCTIONS, EXPRESSED IN AS THOUSANDTHS OF THE MEAN STANDARD DEVIATION OF AGES ELEVEN, TWELVE AND THIRTEEN FOR EACH SEX

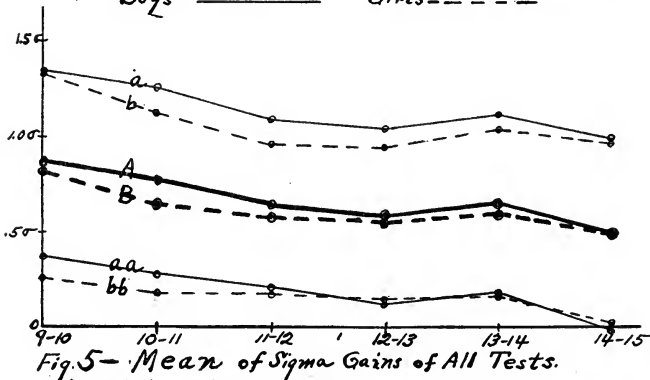
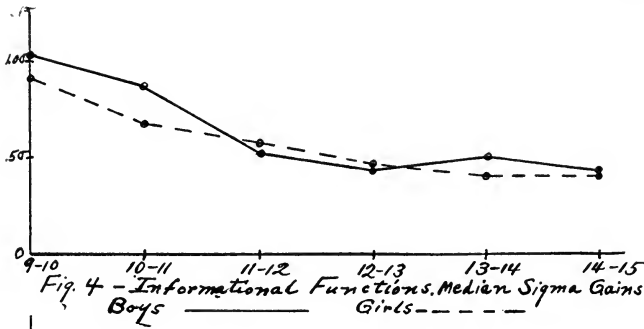
Age	Simple		Memory		Higher		Informa- tional		Combined	
	B	G	B	G	B	G	B	G	B	G
9-10	615	745	770	802	1013	951	1034	806	872	803
10-11	714	607	580	509	845	760	875	675	773	647
11-12	705	419	532	428	726	650	507	573	643	561
12-13	640	554	294	454	675	589	434	467	570	542
13-14	1019	692	095	479	678	683	502	403	647	596
14-15	539	625	269	392	595	542	439	407	489	487

TABLE II.—Q's (EXPRESSED AS THOUSANDTHS OF THE MEAN S. D.'s) OF THE YEARLY GAINS FOR EACH AGE AND SEX IN EACH GROUP OF SIMILAR FUNCTIONS, AND IN THE COMBINED GROUP

Age	Simpler		Memory		Higher		Informa- tional		Combined	
	B	G	B	G	B	G	B	G	B	G
9-10	465	503	863	1187	398	456	413	333	485	545
10-11	509	587	681	652	506	426	324	337	487	470
11-12	500	340	523	645	402	374	391	306	437	391
12-13	448	357	742	529	452	449	314	290	460	403
13-14	519	407	520	603	523	471	302	278	466	430
14-15	548	375	434	583	629	532	311	384	504	471

Figures 1 to 4 show the rates of gain from year to year. Since the average gain per year is very nearly 0.6σ , we have used the same distance to represent 0.6σ in the vertical scale as is used to represent one year in the horizontal scale. Figure 5 shows the rate of improve-





A = Median Gains for Boys

a = 75 Percentile Gains for Boys

aa = 25 Percentile Gains for Boys

B = Median Gains for Girls

b = 75 Percentile Gains for Girls

bb = 25 Percentile Gains for Girls

ment for all tests combined, and also shows the 25 percentile and 75 percentile curves for the combination of all tests.

Results.—In the case of the simpler functions tested by number-checking and handwriting tests, Table I and Fig. 1 show but two departures from a constant rate of growth—the drop in the girls' curve at eleven to twelve, and the sudden rise in the boys' curve at thirteen to fourteen. In memory functions (Fig. 2) there is for boys a sudden drop at thirteen to fourteen. In memory, higher and informational functions, and in the composite of all tests the curves show less gain at the later ages than at the earlier ages.

The irregularities of the curves (Figs. 1 and 2) are probably best explained in terms of the reliability, rather than by such theories as pre-adolescent slowing down, or adolescent spurt; in fact, adolescent spurt is sorely taxed to explain the boys' curves at thirteen to fourteen in simpler and memory functions and at the same time explain their curves in higher and informational functions; especially since the groups tested were exactly the same in both cases.

When we combine the results of more tests we get a smoothing of the curves. When we take the curves for each of the tests we find great irregularity. In another connection¹ we have shown that increasing the number of tests and increasing the number of cases has the effect of smoothing the curves. But, other things being equal, this increase in the number of tests, or in number of cases, is the means of securing more adequate measures of the functions for any age group and increases the reliability. It seems wise, therefore, to regard the irregularities in the curves as probably of chance occurrence.

In considering the apparent decrease in rate of growth in memory, higher, and informational functions, the following facts must be kept clearly in mind:

No nine or ten-year-olds below the advanced fourth grade were tested. No children beyond the third year of the junior high school (grade IX) were tested. These limits cut off younger children of less ability and older children of more ability. Dewey, Child, and Ruml,² making a careful selection from three thousand New York City

¹ Brooks, Fowler D.: "Changes in Mental Traits with Age, Determined by Individual Re-tests," Chap. VI. Here are presented data from other investigations, covering fourteen thousand children of ages nine to fifteen, and four thousand of ages sixteen to eighteen.

² Dewey, Child, and Ruml: "Methods and Results of Testing School Children," 1920.

public school children, so as to secure an unselected school population, found that 34 per cent of the nine-year-old boys and 34 per cent of the nine-year-old girls were below grade IV, while 38 per cent to 46 per cent were in the beginning IV. Of the ten-year-old boys 16 per cent were in or below the beginning grade IV, while 26 per cent of the ten-year-old girls were similarly classified. Since we did no testing below the advanced IV, and since the testing was done at the end of the school year, it is seen that the nine- and ten-year-olds represent a more highly selected group than those of later ages. This, of course, is upon the basis of grade reached in a carefully graded and carefully supervised school as an index of ability.

Under these circumstances we would expect greater gains in the earlier years in those functions which we think of as being more closely associated with intellectual ability. The curves for memory, higher and informational functions seem to show this, or rather, are to be interpreted with this fact in mind. Without doubt a few children of superior ability were not tested at the later ages, having already completed the ninth grade.

Then too, quite a number of the older children made such high scores at their first testing that at later re-tests there was no opportunity to show much gain in gross score in some of the tests (*e.g.*, a twelve-year-old spelling fifty-seven of a sixty word test, could not show the same improvement in gross score during the two following years as could a ten-year-old who spelled twenty-six words at his initial testing; nor could a thirteen-year-old, having an initial score of 182 out of a possible 190 in reading vocabulary gain as much as a ten-year-old scoring 128). There are enough of such cases to account for much of the decrease shown in rate in several of the tests. Then, too, to make a gain of two points from a high initial gross score probably represents greater absolute improvement than the same gain in points from a lower initial score, yet we have no way of allowing for this fact in making up Table I. These two considerations (such high initial scores as allow no chance for much improvement in gross score, and the significance of small increases in gross scores from high initial scores) account for much of the decrease shown at the later ages, and, when considered in connection with the effect of selection at the earlier ages, indicate that the rate of mental growth is probably very nearly constant from nine to fifteen for school population.

To analyze the data still further we have combined the gains into groups of functions upon three other bases of classification: (1) as to

presence or absence of very high scores; (2) as to influence of school instruction; (3) as to ability required to make an initial score. The curves for these groupings show the following resemblances: higher functions with (a) tests in which very high scores were absent, (b) tests much influenced by school instruction, and (c) tests in which making an initial score requires much ability; informational functions with (a) tests in which some perfect or very high scores were made, (b) tests little influenced by school instruction, and (c) tests in which making an initial score requires comparatively little ability. Furthermore, these two sets of curves are not at all dissimilar, the classification upon basis of presence or absence of very high scores giving the only curves that show the same differences as Figs. 3 and 4 in the amount of decrease in rate at the later ages.

Significant sex differences in the rate of mental growth are not shown. Combining results for boys and girls of the same ages, we have twenty-four to forty-five cases of each age group.

Space permits but very brief mention of the light our data throw upon the doctrine of compensation. The correlations between the gains in the different groups of functions have been calculated by the Pearson product-moment formula, age and sex having been equalized. These coefficients, uncorrected for attenuation, are positive but low, varying from 0.04 to 0.31. These indicate that gain in one group of functions is not accompanied by loss or decrease in another group, as has often been stated.

It is further seen that mental growth is here found to be at a regular rate from year to year, and not characterized by sudden spurts. Of course, it may well be contended that such spurts do occur but that the annual re-tests do not show them; that, if we had a great number of reliable individual mental growth curves—curves showing growth during a period of years by three- or four-month intervals—we might find spurts. This contention involves at least the three following considerations:

First.—We must not overlook the fact that the doctrine of spurts, is based, not upon extensive experimental evidence, but upon observation and reasoning by analogy from physical development; that experimental evidence seems to indicate regularity rather than irregularity.

Second.—We do not now have, and in the present state of development and educational tests and scales it is doubtful if we can secure such reliable individual mental growth curves as are needed to settle this question. To do so we need tests of a very high degree of reliabil-

ity; for many functions we need enough alternates of demonstrated equivalence that two tests may be given at any one time (say one day to one week apart) and, three or four testings per year may be made without introducing the practice effect that would come from using the same form of a test three or more times a year. We need, also, to be sure that all who are being tested have a certain familiarity with test procedures, so that the improvement due to becoming more used to tests, may be reduced to a minimum. These are special considerations in addition to the usual standard procedures.

Third.—The re-test method may be called in question on account of the limitations due to our present lack of an adequate number of reliable tests. It may be urged that this deficiency may be overcome if, instead of the re-test method, we test many thousands of each age, grouping them by quarter-years of age. Bureaus of research that have a broad program of measurement can supply much valuable data of this kind. Such bureaus, however, by using suitable individual record cards, and by following a consistent program of testing from year to year, will also have valuable data for such studies of individual mental growth as our present (and of course, future) tests and scales enable us to make. The re-test method, despite the limitations placed upon it by our present tests and scales, has certain well-recognized, fundamental advantages, long ago pointed out by Thorndike.

Conclusions.—(1) We have seen that the re-tests of this group of one hundred seventy-one children, ages nine to fifteen, by a large battery of tests, show mental growth to be at a rate probably very nearly constant from year to year. The variations from straight-line development at a constant rate are probably best interpreted as due to selection at the earlier ages, to the presence of numerous high scores at the later ages, and to the small number of cases.

2. No significant sex differences in rate of mental development are found.

3. Regularity, rather than irregularity, seems to characterize mental development as measured at yearly intervals.

4. Increase in ability in one group of functions does not seem to be accompanied by a decrease in ability in some other group.

5. Wherever careful testing is done, individual cumulative records of test results should be kept. Such procedure, if properly standardized, will give an increasingly large amount of very valuable data on the problem of mental development of school children, and will give it with a minimum of extra expenditure of time and money.

VERBAL AND ABSTRACT ELEMENTS IN INTELLIGENCE EXAMINATIONS

JOHN P. HERRING

Bureau of Educational Research, Bloomsburg State Normal School, Bloomsburg,
Pennsylvania

This is a study of relations existing between human intelligence on the one hand and certain definite abilities on the other. These definite abilities are comprised in two somewhat distinct scales; *first*, a scale of situations extending from very concrete to very abstract; and *second*, a scale of situations extending from very non-verbal to very verbal.

The conclusions of such a study as this should relate both to pure and to applied psychology. In human nature, original and acquired taken together, what relations exist among the abilities named? And in practice, for the purposes of prediction and of control, what relations? How may we best estimate intelligence—through samplings of concrete or of abstract work—non-verbal or verbal?

We will call the scales C for concrete and N for non-verbal.

Scale C involves two concepts: *concrete* and *abstract*. *Concrete* is taken to mean: pertaining to things or objects; having no, or little, reference to the investigation of relations between objects; not problematical in character. *Abstract* is taken to mean: pertaining to relations between objects; involving higher intellectual processes, particularly analytic and synthetic thinking; problematical in character. The distinction is not one between working *with* and working *without* material objects; for the same material things may enter equally into responses that are most widely separated in this scale. The Stenquist Mechanical Tests probably involve far less of the concrete and more of the abstract than is obvious.

As illustrative of concrete situations we may take the following: (1) Threading the mazes of the Army Beta test. More difficult and complicated mazes would be more abstract, more problematical; (2) Classifying blue squares and red triangles as in the Dearborn Intelligence Examination. Tests differing from these, *not* by involving closer discriminations of the same sort but by involving similar and more complicated discriminations, would be more abstract; (3) Shingling a roof after the process is well mechanized. Learning to shingle a roof by experimentation without instruction would be more problematical; (4) Street-sweeping.

As further illustration of *abstract* situations we may take these:

filling a hard completion test; solving hard analogies; reading difficult scientific literature; planning and supervising the construction of large bridges; conducting the business affairs of Standard Oil.

Scale N also involves two concepts, *non-verbal* and *verbal*. Verbal is taken to mean: involving the use of words and other symbols such as numbers, mathematical signs of equality and inequality or any subjective imagery taken to represent and to retain during further study earlier outcomes of investigation. Non-verbal is taken to mean: not involving the use of symbols; pertaining to experiences in which the subject deals immediately with its object; not involving representative symbolism.

Verbal or symbolic situations may be illustrated, obviously enough, by verbal completion and analogies tests and arithmetical problems, and somewhat less obviously by teachers' estimates of children's intelligence; non-verbal non-symbolic situations by writing S between two numbers or symbols that are the same, and D between two that differ; by threading mazes; and by learning digit symbol combinations.

It is assumed for these two scales: *first*, that it is false to classify human behavior dualistically as necessarily either concrete or abstract; necessarily either non-verbal or verbal; and *second* that human behavior may be described in these two phases as comprising responses, the distribution of which is *continuous* throughout both scale C and scale N, some responses being, for instance, extremely concrete; some just a little less concrete, and so on, without gap, throughout the gamut to extremely abstract. These assumptions are roughly borne out by a number of subjective judgments which will be presented.

The definitions of the concepts *abstract*, *concrete*, and *verbal* are not regarded as final or even as necessarily valid, but much rather as starting points for the work of classifying the samplings of behavior elicited by a number of group examinations. These examinations were of two classes, selected for two distinct purposes: *first*, a group, seven in number, employed as a weighted composite criterion of intelligence; and *second*, a group, five in number, employed as the means of measuring concrete-abstract abilities and non-verbal-verbal abilities. The first group, the criterion of intelligence, comprised the following with weights as stated: *First*, the Stanford Revision of the Binet Simon Tests, weight 3; *second*, educational age found by averaging the results of the Thorndike Reading Test Alpha 2, the Woody-McCall Arithmetic Test, the Monroe Arithmetical Reasoning

Test, and an Ayres Buckingham Spelling Test, weight 3; *third*, a measure of intelligence based upon the age and grade reached at the time of the study, weight 3; *fourth*, the National Intelligence Tests, Forms A and B, weight 2; *fifth*, the Thorndike Reading Test Alpha 2, weight 1; *sixth*, teachers' estimates of intelligence, weight 1; and *seventh*, the Kelley Trabue Completion Test Alpha, weight 1.

This weighted composite criterion of intelligence comprises a very wide range of activities, involves several modes of observation obtained upon a number of different and widely separated days, and exhibits self-correlations between 0.9 and 0.8 in age groups. The validity of this criterion is assumed. It is just the sort of criterion psychological investigators everywhere assume. Any who do not accept its validity, may, on that account, differ with the conclusions offered.

The second group, used as a measure of the dependent variables, C and N, comprised the following:

The Dearborn Group Intelligence Examination³
The Army Beta Examination
The Pressey Primer Mental Survey
The Indiana Cross-cut Tests
The Thorndike Visual Vocabulary

All the examinations of both groups were divided into short units, each about a page in length, assumed to be homogeneous with respect both to scale C and to scale N. These units were bound into a book of 63 pages, one page per unit, which was submitted to five judges, who classified its content by assigning each unit a position from 1 to 7 upon both scales. These judgments combined two elements: *first*, an averaging of opinion concerning the content of the concepts themselves (abstract, concrete, verbal and non-verbal), and *second*, an averaging of opinion concerning the content of the situations represented in the book of units. It was important to give each judge a ¹⁰ *role* in the determination of the meaning of the concepts as well as in the classification of the test material. Relatively little stress is laid, therefore, upon the rather formal definitions already presented. To one asking for better definition would be given the average findings of the judges themselves. The term "abstract" comes now to mean—having the character of situations like those marked at or above 5 in this book of units. The uniformity of judgment is indicated by the average correlation, judge with judge, of 0.542 ± 0.019 .

A result of this classification is the existence of two subjective scales, by comparison with which any homogeneous portion of test material or of behavior, may be rated for abstractness and for symbolism. Among the less obvious comparisons which are thus made possible are these:

1. Digit symbol tests are distinctly concrete and very non-symbolic. The "symbols" in such tests are the things themselves, not symbols of other things.

2. The Army Beta Maze Test is as concrete as anything found, and (expectedly) as non-verbal as it is concrete.

3. The Beta Cube-counting Test is distinctly an abstract one and, as would be expected, is very non-verbal.

4. The Woody McCall Arithmetic Test is rated 5 in abstractness and about $3\frac{1}{2}$ in verbal quality.

5. Arithmetical problems are rated very abstract.

6. The Stanford Binet Tests average as a whole rather abstract than concrete, rather verbal than non-verbal, but they do *not* occupy extreme position in either respect. In both scales which extend potentially from 1 to 7 they are rated about 5, leaving in both series about 15 out of 63 samples above.

7. The position of the composite criterion of intelligence is just about 5, like the Stanford Binet, upon both scales.

8. The scale of non-verbal—verbal quality runs from the Beta Maze Test at one extreme to the Thorndike Reading Test Alpha 2 at the other. The scale of concreteness—abstractness extends from the Beta Maze at one end to the Analogies of the National Intelligence Tests at the other.

In the field of applied psychology, for the purposes of prediction and control, the following results and conclusions may be tentatively stated.

The raw correlations include those of four grades in a public school and those of three age groups. These are grades IV, V, VI, and VII. They have membership ranging from 23 to 41 per grade and totalling 118. They are extraordinarily homogeneous with respect to mental age, the standard deviations averaging only 4 months, as against perhaps three or four times that spread in many public school grades. This reduction of variation operates to reduce correlations but leaves them comparable. Averaging the several grades, we have correlations as follows:

Intelligence:

With concrete tests.....	0.25
With the middle portion of scale C.....	0.36
With abstract tests.....	0.58
With non-verbals.....	0.27
With the middle of scale N.....	0.31
With the abstract tests.....	0.54

Similar correlations for age groups consisting of averages of ten year olds, eleven year olds and twelve year olds, are as follows:

Intelligence: with

Concrete tests.....	0.47
Middle Scale C.....	0.58
Abstract tests.....	0.84
Non-verbal tests.....	0.49
Middle scale N.....	0.70
Abstract tests.....	0.88

The P.E.'s of the correlations entering into all these averages range from 0.056 to 0.130 and average 0.10. The average r 's themselves are of course much more reliable. Further, these correlations vary almost as do the S.D.'s of the mental ages; the larger the S.D., the higher the r . So close is the correspondence that the average correlation of the r 's with the S.D.'s is 0.906 for raw and 0.886 for corrected coefficients. Increase in the S.D.'s involved in a correlation results in an increase of correlation which is paralleled by no fact or change in human nature; it is a function of increase in the variability of the group measured. With regard to the correlation existing between the two traits, it is wholly spurious. A correlation of 0.4 in a group having a standard variability of 5 months of mental age may mean just the same degree of mutual implication as a correlation of 0.8 in a group having a much higher standard variability. It appears, therefore, that these correlations have a reliability much greater than the customary statistical formula of probable error reveals.

The correlations point in the direction of the following conclusions:

1. Abstract and verbal tests afford better means for the prediction of human intelligence and the control of human situations than do concrete and non-verbal tests.

2. Abstract and verbal tests will make the best material that can

be selected from such tests as those employed in this investigation for the purpose of inclusion in intelligence examinations.

3. The middle mongrel portions of the two scales are to be distinguished by means of their correlations from the concrete and non-verbal portions; they are in general distinguishably better material for intelligence examinations than N1 and C1 and very inferior to N3 and C3.

4. These abstract and verbal tests, besides being correlated more highly with intelligence, have higher coefficients of reliability and are, for this reason, other things equal, better content for intelligence examinations.

It is by no means argued that it is never proper to use concrete and non-verbal tests of intelligence, for there are circumstances in which they are the feasible form of procedure; nor that they should never be included in a battery of tests for literate and intelligent adults, for the data do not permit the conclusion that concrete and non-verbal tests may not properly supplement the others in a battery of tests.

Perhaps the outstanding conclusion is the following: it seems to be the more purely abstract and the more purely verbal tests that afford the closer measures of intelligence.

We come now to examine the corrected correlations between intelligence and the different portions of these two scales.

COEFFICIENTS OF CORRELATION CORRECTED FOR ATTENUATION, r_{MC1}, \dots, r_{MN3} ,
RAW AND CORRECTED
(M means intelligence)

	RAW	CORRECTED
C1.....	0.474	0.800
C2.....	0.578	0.686
C3.....	0.837	0.963
N1.....	0.492	0.738
N2.....	0.697	0.865
N3.....	0.880	0.951

The following interpretation is suggested:

Total human nature and the mutual demands of human beings have become such that intelligence, as it is required for success in contemporaneous human society, comprises largely the ability to deal effectively with situations involving the use of language and of mathematical and other symbols, both subjective and conventional, and also the ability to control situations through the analysis and inter-

pretation of novel and complicated phenomena. It is hard to think of any important posts of social responsibility, from railway surveying to international law, of which this does not seem true. It is the hod-carriers who typically deal with concrete situations. The master architect controls the situations but he does so through such as the hod-carrier, whom he reaches by the way of many and abstract processes involving complex symbolisms.

WHAT IS READING ABILITY?

J. BENSON WYMAN AND MIRIAM WENDLE

Stanford University

Tests of reading ability have been devised—many of them. Their reliability coefficients have been determined and norms have been obtained; but do their reliability coefficients or their norms guarantee that they are good tests of reading ability or tests of reading ability at all? A reliability coefficient is only one criterion of a test. It measures the amount of agreement one would expect between an individual's score on one form of a test and his score on any other comparable form of the same test; but is there anything to show that the so-called reading tests do measure reading ability?

The present studies were undertaken to get at a method by which we could determine whether the so-called reading tests *do* measure reading ability. The questions arose: "What is reading ability? What criterion can be used for measuring it?" The first suggestion was to use teachers' estimates of the reading ability of their pupils as the criterion. But, knowing the fallibility of teachers' estimates, an alternative criterion was also used and the plan adopted to obtain it was the following: Two professors and three graduate students conversant with the tests used were asked to assign values to each test indicating the value of the test as a measure of silent reading ability (*i.e.*, they gave their judgments of the values of the tests as *tests of reading*, whereas for the first criterion the teachers gave their judgments of the reading ability of the pupils). These five values were made independently. The five values for each test were then averaged and this average value was taken as the scoring, or weight, for the test as a reading test. Then this alternative criterion for reading ability consisted of the combination of all the tests weighted according to the above average values. Having then these two criteria for reading ability the procedure was as follows:

Reliability Coefficients.—The correlation between a given set of scores in one test and another set similarly obtained on a similar form of the same test was determined. In cases where there were not two comparable forms of a test, the one form was divided into two halves measuring substantially the same thing. These halves were correlated; and then Brown's formula ($\frac{2r}{1+r}$)—where r was the obtained correlation—was applied; and this gave an estimate of the correlation

between two forms of the test. This correlation is the reliability coefficient of the test.

Correlations.—(1) Each test was correlated with every other test by means of Pearson's Product-Moment formula and the probable errors of these correlations were determined from

$$P.E_r = \frac{0.6745(1 - r^2)}{\sqrt{n}} \text{ where } \begin{cases} r = \text{correlation} \\ n = \text{number of cases} \end{cases}$$

2. The average of the independent estimates of the reading ability of the pupils made by the two teachers was taken as the first criterion of reading ability. (Call this Reading Ability T.) The sum, or the average, of the pupils' scores on each test was taken as the test score. Then each test score was correlated with Reading Ability T.

3. The second criterion for reading ability (call this Reading Ability C) consisted of the combination of all the tests weighted according to their values as measures of reading ability—with the modification that, when a test was to be correlated with it, that test was taken out of the criterion. (Suppose, for example that the Thorndike Reading Alpha test were to be correlated with Reading Ability, then the criterion for reading ability would be a combination of all the weighted tests except Thorndike Alpha.) In order to determine this correlation, the following formula¹ was used:

$$*r_1(\sum w_x X) = \frac{\sum r_{1x} w_x \sigma_x}{\sqrt{\sum (w_x \sigma_x)^2 + 2 \sum r_{xy} w_x \sigma_x w_y \sigma_y}}$$

where

$\sum r_{1x} w_x \sigma_x$ = the sum of the correlations (each multiplied by its weight and standard deviation) of the type Thorndike Alpha and Monroe, Thorndike Alpha and Completion Beta.

$\sum r_{xy} w_x \sigma_x w_y \sigma_y$ = the sum of all the intercorrelations each one multiplied by the weights and standard deviations of both the tests correlated.

4. Spearman considered that if there were any errors in the original scores of the pupils due to chance mistakes, they would not compensate one another but would reduce the correlation. He devised the following formula by the appli-

¹Kelley, T. L.: *Bulletin* 27, University of Texas, 1916.

cation of which the obtained correlations would be corrected for this, so that the corrected coefficient measures the extent to which the test would correlate with the criterion if the score of the individual were an accurate one both in the test and in the criterion:

$$\text{Corrected coefficient } R = \frac{r}{\sqrt{r_{13}} \sqrt{r_{24}}}$$

where

r = correlation between test and criterion

r_{13} = reliability coefficient of test

r_{24} = reliability coefficient of criterion

R then is the correlation between a true reading score and a true criterion of reading ability.

5. Then that test is more uniquely a reading test and less a test of any other function which shows the highest R .

Detailed Procedure.—Two studies were made. One was with grade VIII B. pupils where reading tests were the main tests, and the criteria were Reading Ability T and Reading Ability C. (We shall refer to this study as "VIII grade reading study.") The other study was with High School pupils, English tests were used and the criterion of English Ability—call it English Ability E—was the Teachers' grades. (This study we shall refer to as "High School English study.")

Tests Given.—(1) The following tests were given to 36 pupils in grade VIII just before promotion. Two teachers independently ranked the pupils in the order of their ability to read:

Thorndike's Reading Scale Alpha 2.

Monroe's Silent Reading Test II (forms 2 and 3).

Thorndike's Reading Test B—Visual Vocabulary (series x and y).

Kelley-Trabue Completion Exercise Alpha.

Terman Group Test of Mental Ability (form A).

Seven S Spelling Test (list 13).

Woody-McCall Arithmetic (forms 1 and 2).

Compositions (1) ("What I should like to do next Saturday."

(2) "The most exciting ride I ever had.")

2. The following tests were given to 94 pupils of the Senior class of a High School:

Briggs' English Form Test (Beta).

Compositions (2).

Thorndike-McCall Reading Scale (form 1).

Kelley-Trabue Completion Exercise (Beta).

Abbott-Trabue tests of Poetic Appreciation (series x and y).

In this study the teachers' grades in English (English Ability E) for the previous semester were then secured as an objective measure against which to gauge these tests as tests of ability in English. The Terman Group Test of Mental Ability had been given, so that the scores on a reliable intelligence test were at hand for comparison. Since it was thought the correlations of the English tests with an arithmetic test might also bring out interesting relations, the scores on the arithmetic exercise in the Terman Group Test were used.

Reliability Coefficients.—For Thorndike's Visual Vocabulary, Woody-McCall Arithmetic, Monroe Silent Reading, Compositions and Abbott-Trabue tests one form was correlated with the second form.

For Spelling, Completion, Terman Group, Opposites Beta, Terman Arithmetic, Thorndike-McCall and Briggs two halves were obtained and correlated and then Brown's formula was applied.

For teachers' estimates (in the first study) one teacher's marks were correlated with the other teacher's and then Brown's formula was applied. But, since the Teachers' grades (in the High School English study) represented the marks of only one teacher per individual, it was necessary to estimate their reliability. In a study of teachers' estimates, T. L. Kelley ("Educational Guidance," sec. 4, page 15) found consistently low reliability coefficients. Teachers' gradings are probably somewhat more reliable. So the reliability of Teachers' Grades was estimated to be about 0.45.

Thorndike's Alpha 2 Test consists of passages that are to be read and questions on the passages are to be answered. In "Difficulty 7" there are two paragraphs in the passage with four questions to be answered on the first paragraph and three on the second. In "Difficulty 8" there are two paragraphs with four questions on each. In "Difficulty 8 $\frac{2}{3}$ " there is one paragraph with four questions on it; and in "Difficulty 9" there is one paragraph with five questions. It will be seen then that there are two ways in which the test can be divided into two comparable halves. The one way is to divide the test so that unbroken paragraphs and twelve questions are in either half. The other way is to divide the test so that there is the same number of questions on either side but neither side has unbroken paragraphs. The test was divided into two parts, according to the former method, by splitting it into its paragraphs so that the errors in

the first four questions in Difficulty 7, in the second four in Difficulty 8 and in all four in Difficulty $8\frac{2}{3}$ formed one part while the rest of the errors formed the other part. These two parts were then correlated; and the reliability coefficient was obtained by applying Brown's formula to this correlation. The test could be divided into two parts, according to the second method, by taking the first half of the sum of the errors in Difficulty 7, and the second half in each of the Difficulties 8, $8\frac{2}{3}$ and 9 as one part and the remainder of the errors as the other part, and correlating them.

Now, whether an individual can answer Question 2 depends to a certain extent on whether he can answer Question 1, whether he can answer Question 3 depends to a certain extent on whether he can answer Question 2, and so on. This is the same for each paragraph. Let us call this correlation between questions based on the same paragraph ρ and the correlation between questions based on different paragraphs η . If we then call the value obtained by correlating the two parts of Alpha 2 according to the first method of division r_1 and that obtained by the second method r_2 , we can determine the value of ρ and η from the following equations:

$$r_1 = \frac{n^2\eta}{n + Mm(m-1)\rho + n(n-m)\eta}$$

where m = number of terms in a group

M = number of groups

n = Mm

$$r_2 = \frac{n(\rho + \eta)}{1 + (n-1)\rho + n\eta}$$

where n = number of terms in either half.

By solving these equations we find

$$\rho = 0.179$$

$$\eta = 0.060$$

Then ρ being greater than η proves that the correlation between answers on a single paragraph is greater than that between answers on different paragraphs. Correlation η is due to a certain intelligence level (a child) acting upon certain independent tasks whereas ρ is due to this plus a factor (operating much as a chance factor) which aids in answering subsequent questions in a set if the first is answered correctly and which hinders if the first is answered incorrectly. Therefore ρ is spuriously high, due to correlation between errors, as a measure

of the correlation between independent questions. Since this is so, r_2 is spuriously high as a reliability coefficient, because the two halves correlated in obtaining r_2 are not composed of independent exercises. Accordingly r_1 is the correct value for the reliability coefficient for Alpha 2.

1. The following are the reliability coefficients for each test (Grade VIII Reading study):

Terman Group Test.....	0.85 ± 0.03
Seven "S" Spelling.....	0.84 ± 0.03
Teachers' Estimates.....	0.84 ± 0.03
Thorndike's Visual Vocabulary.....	0.79 ± 0.04
Monroe Comprehension.....	0.75 ± 0.05
Woody-McCall Arithmetic.....	0.70 ± 0.06
Monroe Rate.....	0.67 ± 0.06
Thorndike Alpha 2.....	0.53 ± 0.08
Kelley-Traube Completion.....	0.50 ± 0.08
Composition.....	0.25 ± 0.10

These reliability coefficients are measures of reliability based on the same pupils. So any difference in them cannot be charged to differences in range of talent. Hence, as regards reliability alone, we can place the above tests in the order indicated by the coefficients.

Terman Group, Teachers' Estimates and Spelling are the most reliable. Thorndike's Visual Vocabulary, Monroe Comprehension, Woody-McCall Arithmetic and Monroe Rate are satisfactory; but neither Alpha 2 nor Completion can be regarded as altogether satisfactory. The reliability coefficient for Composition, based on only two compositions, is very low. If compositions are to be used, by applying Brown's formula it can be seen that in order to have a reliability coefficient comparable to the Terman (0.85) it would be necessary to give 17 compositions:

$$\begin{aligned}
 \text{Reliability coefficient} &= \frac{nr}{1 + (n-1)r} \\
 &= 0.85 = \frac{n(0.25)}{1 + (n-1)0.25} \\
 &= n = 17
 \end{aligned}$$

where

n = number of compositions

r = reliability coefficient for 2 compositions = 0.25

2. The following are the reliability coefficients for each test (High School English study):

Terman Group Test.....	0.94 ± 0.01
Opposites (Beta).....	0.80 ± 0.025
Briggs' Four Test.....	0.78 ± 0.03
Completion (Beta).....	0.78 ± 0.03
Terman Arithmetic.....	0.74 ± 0.03
Thorndike-McCall Reading.....	0.63 ± 0.04
Teachers' Grades.....	0.45 (estimated)
Compositions.....	0.43 ± 0.06
Abbott-Trabue.....	0.37 ± 0.06

The low reliability of the Abbott-Trabue Poetic Appreciation test accords with the results described by Abbott and Trabue in the article entitled "A Measure of Ability to Judge Poetry" (*Teachers' College Record*, March, 1921). Therefore it is not possible to measure poetic appreciation by means of this test. The high reliability coefficient of the Terman test agrees with the findings in the first study. The reliability of Teachers' grades, Composition and Trabue is too low to make them satisfactory measures. (See table, page 526.)

(a) No correlations are exceptionally high. As might be expected the correlations of Terman Group with Completion and Terman Group with Opposites are highest, since both Completion and Opposites are used as Intelligence tests in lieu of using Terman Group. The high correlation between the Terman Group test and the Terman Arithmetic is partly due to its being a correlation between Terman and part of itself.

(b) The highest correlations are those between Terman Group and certain English tests rather than between the English tests themselves. It may be that different aspects of English ability may not exist ordinarily in the same individual; but the implication may be that the Terman Group test, because of its greater reliability, is a better measure of English ability than anyone of the English tests.

(c) The correlation between Teachers' grades (English Ability E) and Thorndike-McCall Reading is the highest; Teachers' grades and Terman Group next; Teachers' grade and Completion next, and then

Teachers' grades and Opposites Beta. But none of these correlations is high, indicating

- (a) that these tests do not measure English ability as it is judged by the teachers, or
- (b) that the reliability of teachers' judgments is low, or
- (c) that other factors than English ability or General Intelligence affect English grades.

It seems that the Terman Group test is as indicative of English ability on the criterion of Teachers' grades as the more apparently English tests are.

2. Correlations, in the VIII Grade Reading study, between Reading Ability T and the tests. (The criterion for Reading Ability T was the average of the independent estimates of the pupils' reading ability as made by the two teachers):

Reading Ability T and Terman Group.....	+0.77 ± 0.05
Arithmetic.....	+0.68 ± 0.06
Visual Vocabulary.....	+0.62 ± 0.07
Seven S Spelling.....	+0.59 ± 0.07
Completion.....	+0.58 ± 0.07
Composition.....	+0.54 ± 0.08
Alpha 2.....	+0.51 ± 0.08
Monroe Comprehension.....	+0.49 ± 0.085
Monroe Rate.....	+0.10 ± 0.11
Age.....	-0.63 ± 0.07

These correlations show that what teachers call "reading ability" correlates more highly with what the Terman test measures than with what the so-called reading tests measure. The rate of reading, as measured by the Monroe tests, shows very little correlation with the teachers' estimates of reading ability. Age within a grade has, as would be expected, a negative correlation with reading ability.

Correlations.—(1) Correlations in the High School English study are:

	English Ability E	Completion	Abbott- Tarbue	Opposites Beta	Composition (Needelson)	Composition (Hillegas)	Briggs' Form	Thorndike- McCall	Terman Arithmetic
Completion.....	0.39 ± 0.07								
Abbott-Tarbue.....	0.31 ± 0.06	0.48 ± 0.06							
Opposites Beta.....	0.35 ± 0.06	0.38 ± 0.06	0.58 ± 0.05						
Composition (Hudelson).....	0.03 ± 0.07	0.28 ± 0.06	0.42 ± 0.07	0.39 ± 0.06					
Composition (Hillegas).....	0.11 ± 0.07	0.32 ± 0.06	0.28 ± 0.06	0.29 ± 0.06	0.43 ± 0.06				
Briggs' Form.....	0.17 ± 0.07	0.42 ± 0.06	0.19 ± 0.07	0.33 ± 0.06	0.30 ± 0.06	0.14 ± 0.07			
Thorndike-McCall.....	0.49 ± 0.05	0.51 ± 0.07	0.54 ± 0.05	0.56 ± 0.05	0.52 ± 0.05	0.29 ± 0.06	0.32 ± 0.06		
Terman Arithmetic.....	0.16 ± 0.07	0.32 ± 0.07	0.15 ± 0.07	0.36 ± 0.06	0.27 ± 0.07	0.14 ± 0.07	0.04 ± 0.07	0.35 ± 0.06	
Terman Group.....	0.42 ± 0.06	0.61 ± 0.05	0.49 ± 0.06	0.64 ± 0.04	0.38 ± 0.06	0.36 ± 0.06	0.38 ± 0.06	0.56 ± 0.05	0.61 ± 0.06

Certain tendencies seem evident here.

3. Intercorrelations between tests (in the Grade VIII Reading study):

	Terman Group	Visual vocabulary	Monroe comprehension	Completion	Thorndike Alpha 2	Composition	Seven S spelling	Woody-McCall Arithmetic
Visual Vocabulary..	0.69							
Monroe Compre- hension.....	0.65	0.44						
Kelley-Trabue com- pletion.....	0.64	0.62	0.30					
Thorndike Alpha 2.	0.58	0.56	0.20	0.54				
Composition.....	0.55	0.45	0.20	0.32	0.57			
Seven S Spelling...	0.53	0.55	0.54	0.21	0.17	0.38		
Woody-McCall Arithmetic.....	0.53	0.39	0.45	0.50	0.24	0.40	0.42	
Monroe Rate.....	0.26	0.23	0.62	-0.10	0.07	0.01	0.80	-0.15

4. Correlations, in the VIII Grade Reading study, between Reading Ability C and the tests:

Weighting of Tests.—The following are the average values or weights determined as described previously:

Thorndike Alpha 2.....	10
Kelley-Trabue Completion.....	8
Terman Group Test.....	7
Visual Vocabulary.....	6
Woody-McCall Arithmetic.....	5
Monroe Comprehension.....	5
Composition.....	4
Seven S Spelling.....	3
Monroe Rate.....	2

Correlations

Reading Ability C and Terman Group.....	0.85
Visual Vocabulary.....	0.76
Completion.....	0.64
Alpha 2.....	0.58
Composition.....	0.57
Monroe Comprehension.....	0.53
Arithmetic.....	0.52
Spelling.....	0.49
Monroe Rate.....	0.16

Here the highest correlation is with the Terman Group test; and, as the correlation is a high one, we must conclude they measure very much the same thing. The question arises, is our criterion for reading

ability really a measure of general intelligence, or does the Terman Group test measure reading ability?

Comparing the correlations 2 and these correlations, we see that, according to either criterion, the Terman Group test ranks highest as a measure of reading ability, just as the Monroe Rate of Reading test shows least relationship. The values for the other tests vary in the two lists. It would seem that the second of the two criteria (Reading Ability C) was the more reliable measure of reading ability, for the teachers' estimates of the reading ability of their pupils are more likely to be tempered by their knowledge of the general intelligence of the individuals.

The best test then for reading ability, as far as our data are concerned, is the Terman. Visual Vocabulary and then Completion and Thorndike Alpha 2 are the next. Rate of Reading cannot be considered a test of reading at all in so far as our criteria measure reading ability.

5. Corrected Coefficients of Correlation:

(a) English Ability E and Thorndike McCall Reading.....		0.92
Abbott-Trabue Poetic Appreciation..		0.76
Kelley-Trabue Completion Beta.....		0.67
Terman Group Test of Mental Ability		0.65
Opposites Beta.....		0.59
Briggs' Form.....		0.29
Terman Arithmetic.....		0.28
Composition (2)		0.26
Composition (1).....		0.07
(b) Reading Ability T and Composition.....		1.29 ± 0.25
Terman Group.....		0.98 ± 0.05
Completion.....		0.98 ± 0.11
Arithmetic.....		0.96 ± 0.07
Visual Vocabulary.....		0.83 ± 0.08
Alpha 2.....		0.83 ± 0.12
Spelling.....		0.77 ± 0.09
Comprehension.....		0.67 ± 0.11
Rate.....		0.15 ± 0.17
(c) Reading Ability C and Composition.....		1.14 ± 0.21
Terman Group.....		0.92 ± 0.03
Completion.....		0.90 ± 0.08
Visual Vocabulary.....		0.85 ± 0.05
Alpha 2.....		0.80 ± 0.09
Arithmetic.....		0.62 ± 0.09
Comprehension.....		0.61 ± 0.09
Spelling.....		0.53 ± 0.09
Rate.....		0.20 ± 0.14

For this group of correlations (*c*) where the criterion was Reading Ability C, the probable errors given are the values when the reliability coefficient of Reading Ability is assumed to be 1. Were it less than 1, the corrected coefficients would be $\frac{1}{\sqrt{r_{24}}}$ times greater and the probable errors would be greater.

In determining the probable errors for these corrected coefficients the following formula¹ was used:

$$P.E. = 0.6745 \sqrt{\frac{R^2}{N} \left[\frac{(1-r^2)^2}{r^2} + \frac{(1-r_{13}^2)^2}{4r_{13}^2} + \frac{(1-r_{24}^2)^2}{4r_{24}^2} + \frac{(1-r_{13})(2-2r^2+r_{13}-r_{13}^2)}{2r_{13}} - \frac{(1-r_{24})(2-2r^2+r_{24}-r_{24}^2)}{2r_{24}} + \frac{r^2(1-r_{13})(1-r_{24})}{r_{13}r_{24}} \right]}$$

Spearman found some corrected coefficients greater than unity. He says, "At most, the corrected coefficient is only the true coefficient plus the error due to testing a limited sample—the general magnitude of such an error is indicated by the so-called probable error; and though a true coefficient cannot exceed unity there is no reason why a coefficient plus an error should not do so. In such a case the coefficient must be taken as 1—this being its most probable value."

The only corrected coefficients that would support Spearman's contention that these coefficients are 1 would be those near 1.00 and having small probable errors.

Suppose a corrected correlation $1 + a$. If its probable error be equal to, or less than, $\frac{a}{3}$ then the fallacy of Spearman's contention (that the most probable value of the coefficient is 1.00) is very evident. It proves that his hypotheses (lack of correlation between errors) are unsound. Suppose, for example, that we have a population of 3600, and the corrected correlation between Reading Ability and Composition is 1.29 ± 0.025 . Spearman would say the true correlation was 1.00—without any regard to the probable error value. This correlation (1.00) is about as unreasonable a value as could be chosen, for the chances that the correlation 1.29 would ever be 1.00 are, in the light of its probable error, infinitely remote.

¹ This formula for the P.E. of a coefficient of correlation corrected for attenuation was derived by Dr. Truman L. Kelley, but has not hitherto appeared in print.

Had Spearman known the probable errors of his corrected coefficients, he would have seen the absurdity of claiming that coefficients well above 1.00 supported his argument that all corrected coefficients tended toward 1.00.

GENERAL CONCLUSIONS REGARDING THE TESTS

1. Certain of the English tests are too unreliable to be worth much in the classification of pupils (*e.g.*, Abbott-Trabue Poetry Appreciation test). Opposites Beta is more reliable, Briggs' Form test and Completion Beta are slightly less satisfactory and the Thorndike McCall still less. None of the English tests has as high reliability as the Terman Group test.

2. As far as raw correlations with English ability are concerned, using Teachers' grades in English as the criterion, none of the correlations for any of the tests is marked, the highest being 0.49.

3. The arithmetic test was included in the battery in the second study to see whether the criterion of English ability and the treatment would result in the Arithmetic test falling into a low position as an English test which would be expected on the *à priori* assumption that English and arithmetic are different functions. Since this corrected coefficient is so low (0.293) it affords objective evidence that arithmetic and English ability constitute two separate capacities.

4. From the point of view of the classification of pupils, it is probable better results can be obtained on the basis of these tests in the second study (except Terman Arithmetic, Briggs' Form and Abbott-Trabue Poetic Appreciation) than can be obtained from Teachers' Grades. Opposites Beta, a 15 minute-test, differentiates more correctly than Teachers' grades and could be used very profitably.

5. If compositions are to be used as measuring ability, the average score on from 15 to 20 compositions must be taken.

6. According to our criteria for reading ability, the Terman Group test of Mental Ability is a better measure of reading ability than any of the other tests used.

7. Of the so-called Reading tests used, the best of them as a test of reading ability is Thorndike's Visual Vocabulary, while Monroe's Rate of Silent Reading test shows almost no correlation with our criteria for reading ability.

8. From these studies can be seen the necessity for having other objective information about a test than its reliability coefficient before the function it measures can be stated.

To these conclusions might be added a warning with regard to the use of Spearman's formula—Care must be taken

- (a) that the halves of the tests used in determining the reliability coefficients be strictly independent and comparable samplings of ability. The tests must be given under similar conditions such as will not lead to spurious correlations.
- (b) that the probable errors be determined, and the results be interpreted in the light of the probable errors.

TERMAN VOCABULARY AS A GROUP TEST

ANGELINA L. WEEKS

Miss Hall's School for Girls

The Terman vocabulary test may so easily be used as a group test that an attempt has been made to measure the reliability of some results obtained in this way. For this purpose, the test was given individually and as a group test to the same pupils in two private schools; one, a girls' school of secondary grade, the other, a grade school.

In the secondary school the time limit method was employed, ten minutes being allowed for fifty words. Each subject was supplied with a pencil, paper, and a type-written copy of the words in one column of the list which appears on the last page of the Record Booklet for the Stanford Revision of the Binet-Simon tests. The instructions were: "Define very briefly the words in this list. It is not necessary to give a full definition like that in a dictionary, but a single meaning is sufficient."

The list of words were so distributed that one half of the pupils used the first column, beginning with "gown," which is referred to in this report as "Series A;" the other half used "Series B," beginning with "orange."

After these written group tests were completed, individual tests were given, according to Terman's directions, to the same subjects. In each oral test, the subject was given the series of words which she did not see in the written test so that no error should arise from a possible difference in difficulty of the lists.

There were fifty-seven girls in the secondary school group examined, all pupils in Miss Hall's School for Girls. The ages ranged from thirteen to seventeen years, with the average age sixteen years. Their ratings in Otis and Alpha tests indicated that the group was above the average grade of intelligence.

The accompanying table gives the averages of results obtained in both written and oral tests, and also the averages obtained from the two word-lists.

The variation in method seems to affect the average very little more than the variation in word-lists. The apparently greater steadiness in the written test, which is suggested by a smaller Q, may be due to greater accuracy in rating the definitions. Some subjects gave them so rapidly that they were not taken down verbatim in oral tests.

NUMBER OF WORDS CORRECTLY DEFINED

Test	Average	Q
Written.....	33.53	4.0
Oral.....	35.56	6.0
Series A ("gown" &c).....	33.55	4.0
Series B ("orange" &c).....	35.54	5.5

The vocabulary indices obtained in group and individual tests were converted into vocabulary ages by the norms of Terman, given in "The Measurement" of Intelligence, and those of Hollingworth, given in "Vocational Psychology." The relative changes in age ratings which accompanied the variation in method of testing are given in the following table.

VOCABULARY AGES COMPARED

TERMAN NORMS

Variation of written index from oral index in units of age groups

Written index	Same	Lower		Higher		Totals
No. of groups.....	0	1	2	1	2	
No. of cases.....	33	18	2	3	1	57
Percentile frequency.....	58	32	3	5	2	100

HOLLINGWORTH NORMS

Written index	Same	Lower	Higher		Totals
No. of groups.....	0	1	1	2	
No. of cases.....	41	13	2	1	57
Percentile frequency.....	72	23	3	2	100

According to the Terman norms, thirty-three cases, or fifty-eight per cent, are in the same age group in both tests; twenty-one, or thirty-seven per cent, are one age group lower or higher in the written test than in the oral test. According to the Hollingworth norms, forty-one cases, or seventy-two per cent, are in the same age group in both tests; fifteen cases, or twenty-six per cent, are one group lower or higher in the written test than in the oral test.

With these subjects, whenever the ages indicated by the group test and individual test were unlike, the age given by the written test

tended to approach more nearly to the chronological age. In ninety per cent of these cases, the group test was as accurate an age measure as the individual test.

The correlation of the results obtained in group and individual tests was $+0.7487$, which, though not very high, is sufficient to warrant the use of the group method of vocabulary testing in secondary schools. By this means, extreme cases can be selected early in the school year.

The results of the vocabulary tests were correlated with those of other tests given to the same pupils in the same year. In these computations and all similar ones reported in this article the formula based on rank order, as found in Thorndike's *Mental and Social Measurements*, was used. The following table presents these coefficients.

CORRELATION OF VOCABULARY WITH OTHER TESTS

	No. of cases	Group test	Oral test
Oral vocabulary.....	57	$+0.7487$	
Word-building.....	53	$+0.497$	$+0.515$
Alpha.....	52	$+0.705$	$+0.750$
Otis.....	53	$+0.700$	$+0.579$
Hard directions Woodworth-Wells.....	51	$+0.401$	$+0.479$
Completion A Trabue.....	46	$+0.527$	$+0.551$
School grades English examination....	51	$+0.721$	$+0.614$
Oral English.....	51	$+0.784$	$+0.662$
French examination.....	51	$+0.325$	$+0.396$
Oral French.....	51	$+0.434$	$+0.617$

In order to study the group method with younger pupils the vocabulary was given to thirty-five elementary children. This group included boys and girls of Miss Mill's School, whose grade distribution was as follows:

Grade	Number of pupils
3	9
4	14
5	2
6	1
7	3
8	6

With the printed list of words, each child was provided with a large sheet of paper on which numbers were written corresponding to the numbering of the word-lists. Series B was used in the written test and series A, in the oral. This was done to remove all chance for the words used in the oral tests to be discussed by the children. The lists are so nearly equal in difficulty that it seemed fair to do so.

The instructions were: "You can see a number before each printed word. On your large sheet of paper you find the same numbers. When I read the word numbered one, you may write the meaning of that word after the number one on your paper. It is not necessary to copy the word. If you do not know a word, leave that line blank. Do not write anything in that space. One short meaning is enough. Spelling does not count in this test. I want to find out how many words you know. Ready. Number one. What is an orange?"

The test was conducted in this way until few or no pupils were writing. The examiner then said: "The last words in this column are really meant for grown people, but, if you know one of them, it will add so much to your credit. I will read the words to you. If you hear one that you can write a meaning for, raise your hand and we will wait for you to write it."

Grades five, six, seven, and eight, who were tested in one group, completed the list of fifty words in fourteen minutes. Few defined words after the thirty-fourth. In grades three and four, only twenty-five words were completed in twenty-four minutes. These younger pupils did not refer to the printed lists, but depended upon the spoken word.

During the week following the written tests, each child was given an oral individual test, exactly in accordance with Terman's directions. Correct definitions were somewhat more frequent than in the written test, but the difference was so small that it would make very little change in the pupils' ratings. In only three cases, and those of pupils below the fifth grade, was there a marked disagreement between the written and oral tests.

The chronological age range of this group was from seven to sixteen years. The average age was 9.8 years, with a median of nine years. The median vocabulary age was ten years. The figures for this group are as follows:

VOCABULARY INDICES

	Oral test	Written test
Average.....	42.5	37.0
Median.....	31.0	34.0
Q.....	11.5	12.5
Correlation		
Oral.....	+0.899
School standing.....	+0.236	+0.345

The relative vocabulary age ratings of these younger pupils in the two tests are shown by the following table.

VOCABULARY AGES COMPARED

TERMAN NORMS

Variation of written index from oral index in units of age groups.

Written index	Same	Lower		Totals
No. of groups.....	0	1	2	
No. of cases.....	18	13	4	35
Percentile frequency.....	51	37	12	100

HOLLINGWORTH NORMS

Written index	Same	Lower					Higher	Totals
No. of groups.....	0	1	2	3	4	1		
No. of cases.....	14	13	3	3	1	1		35
Percentile frequency.	40	37	9	8	3	3		100

When the vocabulary ages obtained in the tests were compared with the chronological ages, the written test was found to be as accurate a measure of chronological age as the oral test in all cases except one. This was an eighth grade pupil who was nearly sixteen years old. The written index pointed to an age of less than thirteen, while the oral index gave an age above fourteen according to the Terman scale and above thirteen according to the Hollingworth scale. In this case, the written test agreed with the school record.

Measures made thus far indicate that the group test method of giving the vocabulary is as reliable as the individual method for children above the fifth grade.

NOTES ON ARTICLES IN EDUCATIONAL PSYCHOLOGY IN CURRENT ISSUES OF OTHER MAGAZINES

REPORTED BY CECILE COLLOTON

Department of Educational Psychology, The Lincoln School of Teachers College

EDUCATIONAL TESTS

Variation of Marking Systems as Diagnosed by Objective Tests. Riverda H. Jordan. Journal of Educational Research, 1921, Oct., 173-179. The distribution of school marks in the 6th, 7th and 8th grades in 10 schools of Minneapolis, showing the widely divergent marking systems. Comparison of marks with scores on intelligence tests.

Results with Standard Chemistry Tests. B. J. Rivett. School Science and Mathematics, 1921, Nov., 720-722. Three Chemistry Tests devised at Northwestern High School, Detroit; (1) symbols of thirty-one important elements, (2) valence of twenty important elements and radicals, (3) twenty formulas of most common compounds. Norms based on results of tests given in all Detroit High schools, Jan., 1921 and June, 1921.

The Conventional Examination in Chemistry and Physics Versus the New Types of Tests. Earl R. Glenn. School Science and Mathematics, 1921, Nov., 746-756. A brief discussion of some preliminary tests with instructions on scoring of tests, and statistical treatment, graphic representation, and interpretation of test scores.

Measuring the Progress of Pupils by Means of Standardized Tests. Samuel S. Brooks. Journal of Educational Research, 1921, Oct., 161-172. How standardized tests are used in the rural schools of Winchester, New Hampshire. Reproductions of individual score cards with scores in graph form showing progress through the year.

INTELLIGENCE TESTS

School Variation in General Intelligence. Warren W. Cox. Journal of Educational Research, 1921, Oct., 187-194. Data on the general intelligence of 24 sixth grades in 24 elementary schools in Cincinnati as shown by Otis Group Intelligence Scale. Study of type of community in which each school is located and correlation of character of community with intelligence levels of pupils.

The Relation of Intelligence to Ability in the "Three R's" in the Case of Retarded Children. Maud A. Merrill. The Pedagogical Seminary, 1921, Sept., 249-274. An investigation of the relation between the intelligence of a group of retarded children and their pedagogical ability as measured by standardized educational tests in reading, writing, arithmetic and spelling.

MISCELLANEOUS

Some Further Studies of Gifted Children. Elizabeth Cleveland. Journal of Educational Research, 1921, Oct. 195-199. Results of studies of three "special advanced" classes compared with a control group of normal pupils in the same schools. Studies include health, nationality, home conditions, types of reading, and recreation, amount of travel, vocational and educational plans.

Motivated Drill Work in Third Grade Arithmetic and Silent Reading. J. H. Hoover. Journal of Educational Research, 1921, Oct., 200-211. Description of certain games and devices utilizing the play instinct in drill work. Results of an experiment with these materials in thirty different third grade rooms, including 571 children in non-drill and 568 in drill sections. Improvements in drill section much more pronounced than in non-drill section.

Comparative Social Traits of Various Races. Charles B. Davenport. School and Society, 1921, Oct. 22, 344-348. A study of racial differences in social traits. Investigation conducted at Washington Irving High School with 51 girls representing ten races.

Who Can Be Educated? Willard W. Beatty. School and Society, 1921, Oct., 311-313. A discussion of the need for a laboratory study of children—say, 60 to 100 children of all races for the period from birth to maturity.

Some Elementary Statistical Considerations in Educational Measurements. J. Crosby Chapman. Journal of Educational Research, 1921, Oct., 212-220. A critique of current methods of obtaining norms of achievement for educational tests; ultra-refined measurement on one side and ignored errors of selection, administration, etc. on the other.

Apperceptive Abilities. Augusta F. Bronner. Psychological Review, 1921, July, 270-279. The discussion of apperception as a mental process. How present tests estimate this ability.

NEW PUBLICATIONS IN EDUCATIONAL PSYCHOLOGY AND RELATED FIELDS OF EDUCATION



1. *Three Books which Deal with Mental Pathology.*—Education has for one of its chief aims that modification of instinctive action, which in current phraseology is called “socialization.” Educators strive to make man over, from what he is by original nature, into what his civilized contemporaries wish him to be. Educational psychology must therefore be concerned with the question. What happens within the organism when an instinctive tendency conflicts irreconcilably with another instinctive tendency, with an idea, with a habit, or with a circumstance?

The three volumes here considered try to answer this question. Within the scope of a brief review it is impossible to do full justice to the discussions, which cover hundreds of pages, but the more interesting points may be outlined, taking the books in what seems to the reviewer to be their order of importance for the study of human behavior.

Rivers’ volume¹ originated in his observations upon the psychoneuroses among soldiers in the great war. It is pointed out that in soldiers the danger-avoiding impulses, normally active, are brought into sharp conflict with ideas of patriotic duty, habits of military drill, and the circumstance of being impressed into military service. In this conflict the danger-avoiding impulses are inhibited from biologically appropriate motor expression, but they do not thereby die out from disuse. The direct motor response being suppressed, is transformed into whatever indirect response will allay the impulses. Thus develop hysterical blindness, deafness, paralysis, and other functional disorders, which enable a soldier to be safe, and at the same time patriotic, obedient and dutiful.

Rivers differs from the other authors to be considered in this review in finding the primary source of hysterical behavior not in sex, but in thwarted “danger-instincts.” He shows how the facts of

¹ Rivers, W. H. R.: “Instinct and the Unconscious.” *Cambridge University Press*, 1920, pp. 247.

civil as well as of military life are to be harmonized with this view. For instance, in civil life women so preponderate among hysterics that the disorder takes its name from the Greek word which means uterus. But men develop hysterical symptoms readily enough in time of war. In civil life women are constantly exposed to the dangers of child-bearing, which are analogous to the sufferings and perils of war, but which men never have to face. When men face war, they face danger, as women face danger all the time; and then hysteria appears among men, also.

In general Rivers finds Freud's doctrines confirmed as regards the great importance of instincts, the unconscious conflict and repression of instincts, and the mental mechanisms resulting therefrom.

Kempf,¹ also, finds that instinctive tendencies do not disappear through training, but, when appropriate and direct motor response is made impossible, eventuate in abnormal behavior. Kempf refers the disorders chiefly to the instinct of sex. Particularly does he lay stress upon the theory of sexual attachment to a parent, either of the same sex or of the opposite sex. As Rivers is able to reconcile the phenomena of mental pathology with violated danger-impulses, so Kempf is able to trace them to sexual impulses. Everywhere Kempf asks whether it is not possible to see a sexual symbol in object, act or word, and finds the answer to be affirmative. Let us rather put the question thus: Is it possible to find an object, act or word which cannot be interpreted as a sexual symbol by one "set" for that percept?

Kempf cites case after case of abnormal behavior, seen chiefly at St. Elizabeth's Hospital, where recovery followed psycho-analysis, conducted from his point of view. As has often been said, however, this is no proof of the correctness of the hypothesis, as the symptoms in these cases also disappear *without* psycho-analysis (the characteristic recovery of the manic-depressive): that the liability to recurrence is less after psycho-analysis is not established by anything Kempf presents.

Although the author's particular stand-point remains thus in question, it is true that he performs a service in emphasizing again the importance for education of the autonomic nervous system, too frequently neglected in educational psychology. In the training of teachers, the central nervous system is stressed, because learning

¹ Kempf, E. J.: "Psychopathology." C. V. Mosby Co., St. Louis, 1920, pp. 762.

subject-matter has chiefly to do with cortical neurones. It is clear that if there are neurone-patterns which cannot be modified by instruction, the fact is fully as important as that there are patterns which can be so modified. If the "sets" of the organism, which originate in the autonomic system, cannot be changed, but can only be suppressed in action by training, then it is essential to know what account should be taken of them in education.

It is doubtful whether the third of these books¹ may properly claim space in a scientific periodical. The author's style and intention appear to be journalistic rather than scientific. It is not clear as to what the author's training has been, which might qualify him to undertake a treatise on the subject considered. It is fairly certain that he has not studied psychology systematically, for otherwise he would scarcely attribute to "academic psychologists" the remarks and doctrines which he does attribute to them, (assuming that he means by "academic psychologists" teachers of psychology in universities and colleges). Inspection of the recommended bibliography confirms this impression. It is improbable that he has devoted himself to biology, for he believes that heredity is a matter of small importance in the study of behavior, thus aligning himself with naive majority opinion, as opposed to expert opinion. "Insanity, feeble-mindedness or criminality are not inherited characters. They are often acquired through either imitation or suggestion, or both" (p. 120). "Most of our heredity is pseudo-heredity which, being simply the shaping influence of our environment, can be defeated as soon as we realize that it is not working for our welfare" (p. 126). These sentences may serve to indicate the author's back-ground in biology.

One who appears to be untutored in psychology and in biology will scarcely command respect as an exponent of scientific thought about human conduct. There seems to the present reviewer to be no reason for commenting further upon this book.

As the study of human nature progresses, it becomes more certain that educators cannot "eradicate" instincts, as formerly it was thought that they might. It becomes more and more questionable as to whether education can even modify the inborn impulses of man, much less eradicate them. Education does succeed in modifying

¹ Tridon, A.: "Psycho-Analysis and Behavior." Knopf, New York, 1920. pp. 354.

motor response; but what is the expense account attached to the modification? Is all the inhibition of instinctive action, which is secured by education, pure gain? Or is there a heavy wastage of abnormal conduct, eventuating as a compromise between the persistent original impulse and the idea or habit that has been taught? If so, is this waste avoidable through improved methods of education? Books like those of Rivers and Kempf have value for educators, because they stimulate thought about these questions, though from their disagreement on fundamental issues they make clear that final answers cannot yet be given.

It is remarkable that in a constant perusal of the psycho-analytic literature now accumulated and current, reference is never seen to the laws of learning, which have been established by the laboratory study of animals. The volumes here reviewed make no reference to this experimental work; yet to the educational psychologist the laws of animal learning seem adequate to include the phenomena of abnormal human behavior. Thorndike discovered how a hungry monkey learns to rush to the top of his cage, when food is placed at the bottom; how to teach a kitten to scratch itself immediately, when restrained behind confining bars. Have these discoveries no meaning for those who write about psycho-analysis? Or do they never come in contact with the literature of experimental psychology? In the case of Rivers, at least, one feels constrained to assume the first of these alternatives.

LETA S. HOLLINGWORTH.





NON-CIRCULATING

Journal of Educational Psychology.
Vol. 12, 1921.

NON-CIRCULATING

